A NEW FRAMEWORK FOR QUANTITATIVE MASS SPECTROMETRY DATA ANALYSIS AND ITS APPLICATION TO DISCOVERY BASED PROTEOMICS IN C. ELEGANS

by

Christopher J. Mitchell

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland

December, 2015

Abstract

My research initially began with the goal of identifying substrates of the *C. elegans* protein tyrosine phosphatase *ptp-3* through the use of quantitative mass spectrometry. To identify substrates of *ptp-3*, we employed stable isotope labeling by amino acids in cell culture (SILAC) labeling of *C. elegans* to detect changes in the phosphorylation levels of a *C. elegans ptp-3* knockout strain as compared to wild type. This strategy was employed because we rationalized that by removing a phosphatase, the phosphorylation levels of its substrates would increase. Therefore, phosphopeptides that are more abundant in the mutant can be inferred as either direct or indirect substrates of the phosphatase.

However, after initial experiments it was discovered that *C. elegans* metabolized one of the SILAC labels, arginine, into other amino acids, resulting in a dataset that could not be accurately analyzed by existing applications. Thus, to salvage the data I built a program, PyQuant, that was capable of processing the mass spectrometry data and capable of accurately quantifying the SILAC data from *C. elegans.* Through the use of PyQuant, over 60,000 phosphopeptides were quantified, as compared to ~12,000 quantified by a another program, Proteome Discoverer.

Following this, we realized that PyQuant could be easily adapted to address many shortcomings of quantitative mass spectrometry data analysis. This work focuses on the development of PyQuant and how it drastically increases the proportion of quantified mass spectrometry data and enables advanced data analysis. Lastly, we show how

PyQuant was capable of handling the metabolic conversion of SILAC labels in *C. elegans* and was used to identify 225 putative substrates of a tyrosine phosphatase.

Advisor: Akhilesh Pandey, M.D., Ph.D.

Reader: Mark Halushka, M.D., Ph.D.

Acknowledgements

I am grateful for the opportunity to complete my PhD under the guidance of Akhilesh Pandey, especially given my presence in the lab was very fortuitous. During my rotation year, two labs I wished to join did not have the funding to take me on. This led me away from the path of basic research and set my search off towards a lab that would poise me to be in a position where I could easily enter both academia as well as industry following my PhD. When I saw Akhilesh's publication record and the astounding breadth of technologies he employed, I knew it would be a great opportunity. Fortunately, Akhilesh felt similarly about my skill set and past experience. Throughout my time here, Akhilesh emphasized the importance of understanding the data and the genesis of it. This is often a grueling, thankless, and tiresome job but its importance cannot be understated and is perhaps the single most important lesson I have learned from my time here.

I would also like to thank my thesis committee members, Hui Zhang, Takanari Inoue, Jiou Wang, and Akhilesh Pandey for their help and insight the past several years during this project. This project began as a study that sought to profile signaling in *C. elegans* but evolved into a computational project with an application in *C. elegans*. For their willingness to adapt and understand the applications of PyQuant, I am very thankful.

I wish to thank Marc Halushka and Alex Baras as well for giving me the opportunity to work on and eventually become a co-first author on miRge. Marc taught me an important lesson that your intuition can be a powerful tool when you are working with data you do not understand. Marc is not a coder, but was able to find numerous bugs

and inconsistencies in miRge solely by using intuition to how data should appear and how he expected it to be modeled. I hope to someday be able to intuit my way around problems like he does. Additionally, Marc served as my reader for this dissertation, which I will be forever thankful for.

My time in the Pandey lab was greatly enhanced by my fellow lab members, which I'm grateful for their support and insights during my tenure. Two people in particular, Derese Getnet and Min-Sik Kim were invaluable resources and friends along the way. Derese handled the insane complexity of immunology for me and was always available. Min-Sik has a truly terrifying knowledge of mass spectrometry and analytical chemistry and never was too short on time to answer my questions.

Lastly, I wish to thank my friends and family for their support the past several years. I made great friends in Baltimore and could not have finished without their support. In particular, over the past two years Xylena has been steady in her support of my work as well as making me a better person. Additionally, she was a great editor of my work and turned my convoluted manuscripts into beautiful, linear stories. My parents, who without any of this would be possible I can never thank enough. From an early age they kindled a curiosity and pride in the work I do – irrespective of whether they understood it.

And here

Table of Contents

List of Tables

Chapter 2

List of Figures

Chapter 2

Chapter 3

CHAPTER 1

Introduction

1.1 Mass Spectrometry and Proteomics

It is no understatement that mass spectrometry is revolutionizing biological studies. Current mass spectrometers are capable of a tremendous throughput, and can survey thousands of proteins from complex sample mixtures. This allows for unbiased, global surveys of biological systems and is the reason mass spectrometry has been a consistent aspect of biomarker and discovery based experiments. Combined with the pace of whole genome sequencing, combined applications of proteomics and genomics have allowed for a comprehensive description of an organism's genome and proteome (Chaerkady et al., 2011; Lasonder et al., 2002). In the context of human disease, this has allowed for identification of differentially regulated structural variants (Wu et al., 2013), annotation of novel coding regions (Kim et al., 2014a), and re-annotation of existing genomic elements such as previously annotated non-coding RNA or pseudogenes (Mitchell et al., 2015).

Before delving into the computational workflows involved in proteomics, a general understanding of how mass spectrometry data is acquired, and relevant terms must be covered. In mass spectrometry, the term MSN (also denoted MS^n) is often used to describe mass spectrum. The 'N' in MSN refers to the degree a sample is isolated for each mass speectra. For instance, a MS1 scan corresponds to the initial, unfragmented mass scan that comes directly from the ionization source. In the MS1 scan, the masses of

unfragmented ion species are measured, which in this manuscript corresponds to peptides. Ions contained in the MS1 scan can then be isolated and fragmented, which produces a MS2 scan. The fragmentation pattern of a MS2 scan can then be used to derive information on the composition of the ion species that was isolated in the MS1 scan. There is no inherent limit to how far a mass spectrometer can isolate and fragment an ion species. For instance, the MS2 can be further isolated and fragmented for MS3 scans, and this process can be repeated iteratively to generate further MSN scans.

Two various modes are used to select ions for fragmentation and generation of MS2 scans: data-independent acquisition (DIA) (Purvine et al., 2003) and data-dependent acquisition (DDA). DIA operates by taking a predefined *m/z* window, fragmenting all ions contained within that window, and then repeating this procedure iteratively until it covers the full *m/z* range of the MS1 scan. DDA, on the other hand, operates by surveying the full *m/z* range of a MS1 scan, and then selects a user-defined number of ions, usually the most abundant ions, for further fragmentation. As DIA is relatively new, DDA is currently the more routinely used method. An important distinction is that DDA is a sampling approach, where a set number of ions are selected, usually the most abundant, and fragmented. In replicate studies, this strategy can lead to inconsistencies within data, as an ion can be chosen for fragmentation is only one replicate. In a complex mixture, where many ions of are similar abundance, this can lead to poor concordance between replicates and hinder the statistical power of an analysis.

Next, we cover what is actually fragmented. In proteomics, there are several methods in which the protein composition of a sample is determined. The most

commonly used method is known as bottom-up proteomics. There are several other methods including top-down, middle-down, and middle-up – but they will not be covered here as they are not frequently used methods. Bottom-up, also referred to as shotgun proteomics, is similar in concept to the idea of shotgun sequencing in genomics where a protein mixture is broken up into fragments, and these fragments are sequenced with the goal of reassembling these fragments into a coherent view of the proteome. The fragmentation of the proteome is accomplished through the use of proteolytic enzymes, such as trypsin, which hydrolyzes protein sequences into predictable peptide fragments. This known fragmentation property is crucial, as it restricts the peptide fragments possible to those obeying the cleavage patterns of each enzyme. Search engines can then use this information to derive a theoretical set of peptides and spectra to match mass spectrum against. Additionally, it restricts the *m/z* range that ion species will be seen at, which allows mass spectrometers to optimize their parameters for a set dynamic range to achieve a higher resolution and mass accuracy.

1.2 Processing of mass spectrometry data in Proteomics

Similar to next-generation sequencing technologies, mass spectrometry is currently in a deluge of data. As the throughput and capabilities of mass spectrometers is rapidly increasing, the computational needs have similarly evolved. Broadly, analysis of mass spectrometry data processing as it pertains to proteomics entails three steps: 1) acquisition of mass spectra, 2) assignment of MS2 mass spectrum to peptide sequences, and 3) inference of proteins from the peptide sequences identified. Optionally, a workflow may attempt to quantify the abundance of a peptide species through various

methods. The first step of data acquisition is performed by a vendor specific software that generates a file annotating the mass spectra. As such, it is outside of the scope of this manuscript.

Following spectra acquisition, MS2 spectrum are assigned to peptide sequences. The task of assigning MS2 spectrum to peptide sequences is carried out by two complementary methods: *de novo* sequencing and database based spectral assignments. *De novo* methods attempt various combinations of peptide sequences and choose the most likely peptide sequence that corresponds to the observed fragmentation spectrum. Thus, there is no prior information assumed about the composition of the sample. Examples of existing *de novo* software platforms are PEAKS (Ma et al., 2003) and Lutefisk (Taylor and Johnson, 1997). An alternative, and more commonly used, approach is the use of protein database driven search engines such as X!Tandem (Craig and Beavis, 2004), Mascot (Perkins et al., 1999a), or SEQUEST (Eng et al., 1994). In this method, a user defines what proteins they wish to identify in a sample and the search engine produces theoretical spectra corresponding to the enzymatically cleaved peptides from the user supplied input. Then, the search engine proceeds to match experimental spectrum against theoretical spectra to determine which theoretical spectra best explain the experimental data. This approach is inherently limited by the user supplied input, but for well annotated species it is a significantly faster and less computationally intensive than *de novo* methods.

Following assignment of spectra to peptides, the task of inferring the proteins each peptide fragment came from begins. Because of similarities within gene families

and across genes, there are many peptide sequences that can map to multiple protein locations. As such, this results in an ambiguity in how to correctly infer which proteins are present in a sample given a list of peptide sequences. Correspondingly, the correct way to perform this is a contentious subject. A common method for assigning peptides to proteins is the application of Occam's Razor, where the most parsimonious and minimal number of proteins given a set of peptides is reported (Serang and Noble, 2012). Though there is technically no correct approach, it is important to keep the goals of the experiment in mind. For instance, in a purely computational study with no additional validation methods, a conservative approach is appropriate. However, if the researchers are willing to perform additional validation through methods such as western blotting, a liberal assignment of peptides to all possible protein species is acceptable.

The above methods are primarily concerned with the identification of peptide and protein species. For many applications, such as identifying the proteins present in a sample or discovery of novel genes, a qualitative analysis of mass spectrometry data is sufficient. However, when quantitative information is utilized, powerful analyses are now possible. For instance, a tumor and a normal tissue can be compared to identify proteins that correlate with a pathology (Slebos et al., 2015), or changes in phosphorylation levels after a cytokine is applied can reveal new members of well established signaling networks (Pinto et al., 2015; Zhong et al., 2012). Thus, quantitative information is a decisive benefit of mass spectrometry. In proteomics, there are two common methods for peptide quantification, which I will group into MS1 and MS2 based methods.

MS1 based quantification relies on information derived from MS1 scans to estimate a peptide's abundance. Broadly, the central idea of MS1 based quantification is that the abundance of an ion species is proportional to the amount of peptide which was eluted and ionized. Thus, by measuring the abundance of an ion species over a time, we establish the elution profile of that ion, which is a means to measure the peptide's abundance. In practice, a peptide is identified by its MS2 fragmentation spectrum, and then the precursor ion in the MS1 scan that produced the MS2 scan is identified. Because a peptide can elute over several minutes, the precursor ion is searched for in neighboring MS1 scans until it can no longer be detected. Taking the intensity of the precursor ion as a function of the time each scan was acquired produces the extracted ion chromatogram (XIC). By integrating the XIC, the abundance of the peptide is determined.

MS1 based quantification is useful because it permits quantification of a sample without the need for any special reagents. However, to enable studies that focus on relative differences between samples, techniques such as stable isotope labeling in cell culture (SILAC) have been developed that introduce an isotopically labeled amino acid into a sample (Ong et al., 2002). This results in a known mass shift between the unlabeled and labeled sample in the MS1 spectra. Thus, the intensity of the unlabeled and labeled pair can be used to provide a relative estimate of a peptide's abundance between two samples. This technique is particularly powerful because it allows the sample to be mixed together following lysis and analyzed on the mass spectrometer simultaneously. Therefore, variables that may arise from processing two samples in parallel are mitigated.

MS2 based quantification, on the other hand, relies on chemically coupling a tag to a peptide, often through the use of reactive amine groups on free N-termini or lysine residues. Upon fragmentation, this tag is fragmented into an ion of a calculated mass, which is referred to as a reporter ion. Through the use of multiple tags of equal mass but differing reporter ion masses, it is possible to compare multiple samples simultaneously in an MS2 spectrum. A crucial property of this method is that every tag prior to fragmentation is of equal mass and therefore appears at the same *m/z* value in MS1 scans. Two common implementations of this method are known as isobaric tag for relative and absolute quantification (iTRAQ) (Ross et al., 2004) and tandem mass tags (TMT) (Thompson et al., 2003). A primary advantage of these methods is that the sample does not have to be grown in the presence of an isotope, which enables its use in human samples that would otherwise be problematic to metabolically label. However, a downside of these methods is that unlike MS1 scan intensities, MS2 based intensities are not comparable to other MS2 scans. Thus, this method enables comparisons between samples of a given peptide, but is not well suited for comparing across peptides of a given sample.

CHAPTER 2

PyQuant: A versatile framework for quantitative mass spectrometry data analysis

Introduction

The utility of quantitative MS data is highly dependent on the accuracy and comprehensiveness of the tools used for analysis. The first step in quantitative proteomic data analysis is to associate spectral information of MS and MS/MS scans with peptide sequences, followed by quantification of the identified peptides. This provides two avenues to increase the rate of quantification: an increase in spectral assignments and/or an increase in the fraction of scans that are quantified. To increase the number of spectral assignments, common approaches include the integration of multiple database search engines, iterative searches, wide-tolerance searches, and custom database searches (Alves et al., 2008; Chick et al., 2015; Kwon et al., 2011; Renuse et al., 2011; Schwartz and Gygi, 2005; Shteynberg et al., 2011; Tharakan et al., 2010; Wang et al., 2012). The major benefit of these approaches is an increase in the number of spectra assigned to peptides. While increasing the number of spectral assignments allows for a greater number of quantitative measurements, it does not address a more fundamental issue that programs used for quantitation only quantify a subset of the available data. Analogous to database search engines, each program used for quantitation has its own method for quantifying data. Thus, data quantified by one program may be missed by another and there may be inherent limitations that are common to several programs.

An emerging issue as well is that the pace of mass spectrometry data acquisition is outstripping many of the existing tools. Currently, the leading platforms for mass spectrometry analysis are MaxQuant (Cox and Mann, 2008) and Proteome Discoverer. Both platforms are tied to the Windows operating system and GUI based. Because of this design, both platforms are unable to make use of computational clusters and require user interaction for every experiment run. These features are inherently at odds with automated high throughput data analysis pipelines. A third computational workflow, the Trans-Proteomic Pipeline (TPP) (Deutsch et al., 2010), is capable of being run via the command line and can be used to create highthroughput workflows. However, its ability to perform quantitative analysis of mass spectrometry data lags significantly behind the capabilities of MaxQuant and Proteome Discoverer and thus, it has not been as well adopted.

To address these issues, we have developed PyQuant as a new framework for quantitative analysis of MS data. PyQuant is a versatile, cross-platform quantitation tool that can be used in conjunction with existing data analysis frameworks or as a quantification node for a minimal, light-weight mass spectrometry data analysis pipeline. PyQuant accepts a variety of input formats, including the mzML and pepXML formats, several vendor specific formats and allows for generic tab delimited input for formats that are not natively supported. Additionally, it is compatible with widely used quantitative MS methods such as metabolic labeling (e.g. SILAC, NeuCode, ^{15}N, and ^{18}O) and chemical tagging (e.g. iTRAQ and TMT) and allows users to define custom labeling and data quantitation strategies. PyQuant can serve as a simple, stand-alone quantitation

program following peptide assignment of many search algorithms such as Comet or X! Tandem (Craig and Beavis, 2004; Eng et al., 2013). This allows the user to have a minimal, easily deployed pipeline for mass spectrometry data analysis. Also, PyQuant can serve as a post-processor of data analyzed with existing frameworks such as MaxQuant, Proteome Discoverer, or the Trans-Proteomic Pipeline (TPP), which allows for an additional, independent algorithm to verify existing quantification values as well as quantifying data excluded by other algorithms.

Materials and Methods

Preparation of peptide samples for standard MS dataset

A lysine and arginine auxotrophic *E. coli* strain, *SLE1*, was purchased from the *Caenorhabditis Genetics Center* (CGC) and grown in M9 Minimal Media (Cold Spring Harbor Protocols). Each culture was grown with a mixture of SILAC isotopes at the desired mixing ratio, lysed, and fractionated via SDS-PAGE. Gel pieces were excised and destained with 40 mM ammonium bicarbonate and 40% acetonitrile (destain buffer) at room temperature. Following destaining, samples were reduced by incubation with 5 mM dithiothreitol (DTT) for 10 minutes at 60^{o} C. After reduction, samples were alkylated by 10 mM iodoacetoamide for 20 minutes. Following reduction and alkylation, samples were washed with destain buffer and incubated with 100% acetonitrile on ice until dehydrated. After dehydration, samples were resuspended in 10 μg/ml trypsin until the samples were rehydrated. After this, excess trypsin was removed and the sample was digested overnight. Elution of digested peptides was performed by adding 80%

acetonitrile with 0.1% trifluoroacetic acid (TFA) and incubated at 25° C on a shaker for 20 minutes. After 20 minutes, the supernatant was removed and the eluted peptides were lyophilized and stored at -20° C until LC-MS/MS analysis.

For labeling of *C. elegans, E. coli* strain *SLE1* was purchased from the CGC and labeled with light and heavy amino acids. Light cultures were grown in LB and heavy cultures were grown in M9 Minimal Media (Cold Spring Harbor Protocols) with the isotopically labeled amino acids Arginine10 and Lysine8. Each culture was grown overnight, and centrifuged at 8,000 x g for 10 minutes at 4°C. For each 250 mL of culture pelleted, 50 mL of S. Media (Stiernagle, 2006) was added and the pellet was resuspended. For culturing of *C. elegans,* 50 mL of resuspended *E. coli* was added to a 1 L flask with *C. elegans* and 10 μg/mL of Nystatin to prevent fungal growth. The culture was monitored and when the bacteria was almost clear (approximately 3 days) and the *C. Elegans* were isolated in accordance with previously described methods (Stiernagle, 2006). *C. elegans* were resuspended in 9M urea containing a protease inhibitor cocktail (Roche #10711400), 50 mM beta-galactosidase, 25 mM sodium fluoride, and 1 mM sodium orthovandate. Next, the resuspension was tip sonicated and viewed under a microscope following sonication until the *C. elegans* were sufficiently broken apart. Following this, the lysate was cleared by centrifugation at 14,000 x g for 20 minutes at 4°C and protein concentration was measured via BCA. 250 μg from light and heavy *C. elegans* were combined for a total of 500 μg. The lysate was then alkylated with 5 mM of iodoacetamide and reduced with 5 mM of dithiothreitol. Following this, the lysate was diluted to a final urea concentration of 3M and digested with Lys-c for 4 hours at room

temperature. After digestion with Lys-c, the lysate was diluted again to a final urea concentration of 1.5M and digested overnight with trypsin. After confirming digestion efficiency, the sample was acidified with 1% TFA and centrifuged at 2000 x *g* for 5 minutes at room temperature. The supernatant was then loaded onto a Sep-Pak cleanup C_{18} column (Waters, Cat#WAT051910) equilibrated with 0.1% TFA. Columns were washed with 12 mL of 0.1% TFA and peptides were eluted with 6 ml of 40% acetonitrile with 0.1% TFA. Eluted peptides were then lyophilized and subjected to basic reverse phase liquid chromatography and then mass spectrometry analysis.

LC-MS/MS

Peptide samples were analyzed on an LTQ-Orbitrap Elite mass spectrometer (Thermo Electron, Bremen, Germany) interfaced with Easy-nLC II nanoflow liquid chromatography systems (Thermo Scientific, Odense, Southern Denmark). The peptide digests from each fraction were reconstituted in Solvent A (0.1% formic acid) and loaded onto a trap column (75 µm × 2 cm) packed in-house with Magic C18 AQ (Michrom Bioresources, Inc., Auburn, CA, USA) (5 µm particle size, pore size 100 Å) at a flow rate of 5 µl/min with solvent A (0.1% formic acid in water). Peptides were resolved on an analytical column (75 µm × 20 cm) at a flow rate of 350 nl min−1 using a linear gradient of 7–30% solvent B (0.1% formic acid in 95% acetonitrile) over 60 min. Mass spectrometry analysis was carried out in a data dependent manner with full scans (350–1,800*m/z*) acquired using an Orbitrap mass analyzer at a mass resolution of 120,000 in Elite at 400 *m/z*. The twenty most intense precursor ions from a survey scan were selected for MS/MS from each duty cycle and detected at a mass resolution of 15,000 at a *m/z* of

400 in the Orbitrap analyzer. All the tandem mass spectra were produced by higher-energy collision dissociation (HCD) method. Dynamic exclusion was set for 30 s with a 10 p.p.m. mass window. The automatic gain control for full FT MS was set to 1 million ions and for FT MS/MS was set to 0.05 million ions with a maximum ion injection times of 100 ms and 200 ms, respectively. Lock-mass from ambient air (*m/z* 445.1200025) was used for the internal calibration.

Data Analysis

Data analysis on the Proteome Discoverer platform (version 2.0) was performed using MASCOT (version 2.2.0) (Perkins et al., 1999a) and SEQUEST (Eng et al., 1994) as the search algorithms. For both MaxQuant and Proteome Discovered, the search parameters allowed for two missed cleavages; carbamidomethylation at cysteine as a fixed modification; N-terminal acetylation, deamidation at asparagine and glutamine, oxidation at methionine, and the appropriate SILAC labeling provided as variable modifications. For *C. elegans* data, the variable modifications of phosphorylation at serine, threonine and tyrosine were specified. MS data was acquired on the LTQ-Orbitrap Elite mass spectrometer, the monoisotopic peptide tolerance was set to 10 ppm and MS/MS tolerance to 0.1 Da. The false discovery rate was set to 1% at the peptide level.

For post-hoc analysis of MaxQuant and Proteome Discoverer, PyQuant was provided with the output of each respective program to perform an additional round of quantification. PyQuant was provided with the msf output of Proteome Discoverer and mzML files of the raw data. For *C. elegans*, the additional parameter of "--spread" was

supplied; otherwise the default parameters were retained (precursor mass error of 5 ppm and isotope-selection error of 2.5 ppm). For MaxQuant, the ms_ms.txt table was provided as input.

PyQuant's methodology and availability

PyQuant is a command line driven program developed using the Python programming language, and is compatible with both Python 2.7+ and Python 3.5+. For GUI-based access, PyQuant can be deployed using the Wooey framework [https://github.com/wooey/Wooey/]. The web based interface uses jQuery, DataTables, pako, and c3 for visualization. Installation instructions, source code, and guides for PyQuant are available at https://pandeylab.github.io/pyquant/ and it is compatible with most major operating systems.

Although PyQuant is capable of quantifying at any ms^n levels and is highly customizable, for simplicity we describe PyQuant's algorithm by stepping through two common quantification scenarios: a traditional proteomics label-free data analysis pipeline, in which the precursor ion of a ms^2 scan is quantified, and an experiment utilizing chemical tags such as iTRAQ, where the ms^2 scan is used to identify and provide quantification of a peptide. However, PyQuant can be configured to quantify at any ms level, which allows it to be used with emerging mass spectrometry strategies such as ms^3-based quantification. Finally, we adopt the term analyte to comprise the physical entity of interest, such as a peptide and we use the term ion species to refer to a distinct m/z value. The analyte may be comprised of one or more ion species, such as the

monoisotopic peak as well as peaks resulting from the inclusion of naturally occurring isotopes.

The steps of PyQuant can be broken down into several distinct steps: input data processing, peak picking, and quantification of the extracted ion chromatogram (XIC) (**Fig. 1**). Because of PyQuant's flexibility, there are many ways peaks can be assigned and selected. Thus, we cover first how PyQuant processes different types of data, and end with a step common to all algorithms, quantification of the XIC.

Input Data Processing

PyQuant accepts a variety of data as input. For data which has been processed with other frameworks, PyQuant currently supports the pepXML files created by the TPP, msf files produced by Proteome Discoverer, and ms_ms text files produced by MaxQuant. For search engines, PyQuant can parse the output of X!Tandem (Craig and Beavis, 2004), search engines that produce pepXML files such as COMET (Eng et al., 2013), and Mascot (Perkins et al., 1999b) if the ms_parser library is installed. To provide compatibility with an assortment of programs, PyQuant can also be supplied with a generic tab delimited file providing information on spectral assignments. Lastly, PyQuant is able to operate solely on raw mass spectrometry data in the mzML format.

Processing of raw data

In the absence of any provided annotation, PyQuant will quantify all ms^1 scans with a corresponding ms^2 scan. This can be used to perform simple analysis of raw data, such as identifying intense peaks that may represent sample contaminants, unspecified

modifications, or interesting biological findings such as novel or non-tryptic peptides that were missing from a database search. PyQuant can also be provided with target ions to search for in raw data, which can be used to identify missing values between replicates or perform targeted searches for post-translational modifications through the use of signature ions.

Processing of mixed resolution data

It is sometimes useful to acquire scans at different resolution levels, such as in NeuCode labeling. To assist with this analysis, PyQuant allows a user to define a minimal resolution power of a scan to be considered for quantification. This allows the user to quantify scans using this time saving data acquisition strategy, which is becoming more popular as the resolving power of mass spectrometers increases.

Processing of peptide spectral assignments

After being provided with the output of a given search engine or platform, PyQuant begins by parsing result files for peptide spectral assignments (PSMs). Following this, PyQuant checks whether a labeling strategy has been defined. For known data formats that define the labeling strategies, such Proteome Discoverer's msf file or X! Tandem's XML output, any user defined labeling strategy is automatically parsed and applied. However, for cases where this information is not embedded in the file, or if a vendor format changes, the labeling strategy can be supplied as a simple tab delimited file. Next, because the mass error changes as a function of the m/z of an ion, the error between the theoretical peptide's mass and the observed mass is plotted and fit to a spline

function (**Supplementary Figure 1**). This spline is used to offset the machine drift with higher m/z values, which would otherwise make it difficult to identify values at a higher m/z within a given error tolerance. PyQuant then identifies the precursor ion of PSMs, and if a labeling strategy has been defined, PyQuant searches for corresponding labeled peaks in the quantification scans irrespective of whether these peaks have been fragmented. This helps to identify and quantify peaks that were not selected for fragmentation.

At this point, PyQuant has identified precursor ions and any applicable labeled peaks from the PSMs. If the user chooses to quantify scans at the ms^1 level, PyQuant identifies the isotopic cluster of each precursor ion. However, because each isotopic cluster may be comprised of a mixture of overlapping ion species, the profile for each ion is fitted to a Gaussian mixture model to remove any interfering ions (**Supplementary Figure 2**). Following this, the peaks for each isotopic cluster are integrated to provide an abundance measure for each ion species detected. Finally, this process is repeated for neighboring scans until all ion species comprising an analyte can no longer be identified in two subsequent scans.

Quantifying the extracted ion chromatogram

To measure the abundance of a given analyte, the extracted ion chromatogram (XIC) is generated by extracting intensities from each isotopic peak of an ion species and plotting them against their respective retention times. For each XIC, multiple Bi-Gaussian peaks are fit to the data to accurately model the ion species of interest as well as

any contaminating ions. A Bi-Gaussian peak was chosen because it better models the skewness of elution profiles as compared to a Gaussian distribution (**Supplementary Figure 3**). To avoid over-fitting the data with multiple Bi-Gaussian distributions, the number of Bi-Gaussian peaks modeling the data is determined using the Bayesian information criterion (BIC) (Schwarz, 1978), which guards against over-fitting by penalizing additional parameters. Following this, the mean and standard deviations of each peak is compared to the retention time the initial scan was identified at, and peaks not containing the retention time are excluded. This routine is performed for all isotopic clusters and labeled pairs found for a given ion species. Next, because the elution profile for isotopic clusters and their isotopically labeled pairs should be similar, the peak shape of all isotopic clusters are compared and outlier peaks are identified and excluded by the minimum covariance determinant method **Supplementary Figure 4A-C**) (Rousseeuw and Driessen, 1999). Lastly to measure the abundance for each ion, the area under the XIC is integrated over.

Confidence Estimates of Quantification

PyQuant uses several approaches to quantify data. However, this can result in the inclusion of noisy, unreliable data. To guard against this, PyQuant employs several methods to provide an estimate of the accuracy of quantification. PyQuant uses a machine learning algorithm trained with a manually curated dataset of “good” and “poor” fits to provide a confidence measurement of the quality of fits. This takes into account variables such as the signal to noise ratio, the intensity of an analyte, the width of an analyte's retention time, and the density of the fitted peak to assess how accurate a given

measurement is. Within the dataset of known ratios, selecting only high confidence fits results in a tighter distribution of quantified peptides (**Supplementary Figure 5**). Lastly, because there is no substitute for manual validation, PyQuant provides an interactive HTML based output that allows a user to filter values, assign cutoffs, and manually inspect every isotope selected and the XIC.

Data Output and Visualization

PyQuant provides two files for the user. One is a tab delimited file that can be easily viewed in excel or processed further in a given data analysis pipeline. The other is a HTML file that offers an interactive browser-based exploration of the data. PyQuant uses several open source JavaScript libraries to provide an intuitive user interface with advanced data table manipulation. Additionally, PyQuant provides detailed graphics depicting which isotopes were selected in each scan, the XIC with corresponding peak fits, and integration values for every ion species. This allows the user to make an informed choice on whether an ion deemed significant by their chosen criteria, such as a fold change between samples, is truly significant or merely an algorithmic error.

Results

The goal of PyQuant is to provide a scalable, versatile framework for MS-based quantitative analyses. To demonstrate the various analyses PyQuant enables and its novel capabilities, we used publicly available datasets that utilize SILAC (Zhong et al., 2012), NeuCode (Merrill et al., 2014), ^{15}N, iTRAQ (Hultin-Rosenberg et al., 2013), and ms^3 based TMT technologies (Murphy et al., 2015).

PyQuant accurately quantifies a variety of quantitative MS data

To assess PyQuant against a SILAC sample with a known isotopic mixture, three separate *E. coli* cultures were grown with a mixture of light, medium and heavy arginine and lysine amino acids in different ratios (**Table 1**). Following overnight growth, the cultures were lysed and processed for LC-MS/MS as described under materials and methods. The raw data were analyzed with two existing platforms, MaxQuant and Proteome Discoverer in addition to PyQuant. As shown in **Fig. 2A**, PyQuant was able to recapitulate the known sample ratios, and consistently matched SILAC ratios given by Proteome Discoverer and MaxQuant. Thus, PyQuant is able to match these existing algorithms in quantifying data of known ratios.

While SILAC uses labeling of select amino acids, a similar technique, ^{15}N, labels organisms by the metabolic incorporation of ^{15}N isotopes into any nitrogen containing amino acids (Washburn et al., 2002). To evaluate PyQuant's ability to handle ^{15}N-based quantitative MS data, a ^{15}N labeled mouse liver was compared to an unlabeled mouse liver. Equal amounts of proteins extracted from each mouse liver were mixed and processed for LC-MS/MS as described under materials and methods. Following peptide assignment with Proteome Discoverer, the raw data and search results were processed with PyQuant to quantify the relative abundance of ^{15}N to ^{14}N containing peptides. Due to MaxQuant and Proteome Discoverer's inability to quantify ^{15}N data, to evaluate the performance of PyQuant, spectra were manually interrogated and compared to the ratio provided by PyQuant, which confirmed that PyQuant provides values that are consistent with manual interpretation (**Fig. 2B**).

A recently developed *in vivo* labeling technique is NeuCode, where various combinations of ^{13}C, ^{15}N, and ^{2}H are incorporated into an amino acid in order to label a cell or organism (Hebert et al., 2013). Due to the number of carbons, nitogens, and hydrogens in an amino acid, the number of combinations of these isotopes results in marked increase in multiplexing capabilities, with up to 36 different mass combinations for lysine alone. Currently, no available software supports quantification of NeuCode isotopes. Thus, we tested PyQuant's ability to quantify this type of data. An experiment with known mixtures of NeuCode labeled lysates was analyzed with PyQuant (Merrill et al., 2014). As shown in **Fig. 2C**, PyQuant was capable of quantifying NeuCode data and the median quantification values matched the known isotopic mixtures of each NeuCode reagent.

For chemical tagging methods, two common methods are iTRAQ and TMT. Each method uses isobaric tags that do not cause a mass shift that can be detected in ms^1 scans, but upon fragmentation the tags provide relative abundance of each labeled sample. Since both methods enables multiplexed comparisons of samples that may not be metabolically labeled, such as clinical samples, it has been widely employed in the quantitative proteomics field. However, one complication which arises in ms^2 based quantification is interference from co-fragmentation of background ions, which can systematically skew relative abundance measurements (Ow et al., 2009). A solution to this complication is the use of ms^3 for quantification as opposed to ms^2, which reduces the chance of co-fragmentation of unwanted ions from MS/MS (Ting et al., 2011). We assessed PyQuant's ability to quantify at the ms^3 level by comparing values derived from PyQuant to

previously published ms^3 based quantifications (**Fig. 2D**) (Murphy et al., 2015). This revealed that PyQuant was nearly identical to the previously published values and is capable of ms^3 based quantification.

Increased quantification rate by PyQuant

Because different peaks could be picked by different quantitation programs, employing multiple programs should increase the proportion of peptide ions that are quantified. In the *E. coli* data used above, we wished to determine the extent to which PyQuant could increase quantification. As shown in **Fig. 3A**, when analyzing a simple *E. coli* mixture, PyQuant increased the rate of quantification by 68% and 44% when compared to Proteome Discoverer and MaxQuant, respectively. To determine how confident the quantification of these peptides missed by other programs was, we evaluated many of peptides which MaxQuant and Proteome Discoverer failed to quantify in PyQuant's output, and determined most of them were accurate (**Fig. 3B**).

PyQuant can quantify isotopically-labeled peptides with unexpected isotopic patterns

SILAC was initially developed for cell culture systems but is increasingly being used for labeling model organisms for global, quantitative analysis of protein abundance. A potential complication of SILAC in organisms is the metabolic conversion of experimentally introduced labeled amino acids into other amino acids (Borek et al., 2015; Van Hoof et al., 2007). Normally, the incorporation of naturally occurring carbon and nitrogen isotopes results in the spread of a peptide from its monoisotopic mass, with each combination of isotopes appearing as a distinct species in a mass spectrum. Without

amino acid conversion, this pattern can be theoretically calculated and used to identify which peaks correspond to a peptide of interest. However, when there is an additional source of labeled isotopes, such as metabolic conversion of labeled amino acids, the isotopic cluster widens and deviates significantly from the theoretical distribution, thus complicating the interpretation of protein abundance.

To evaluate how isotopic conversions impact the quantification of existing programs, we labeled the nematode *C. elegans* with heavy and light amino acids by growing them in the presence of labeled bacteria. As shown in **Fig. 4A**, the use of arginine as an isotopic label resulted in the spread of the heavy isotopic cluster, which deviated significantly from the isotopic distribution of the unlabeled peak. This data was quantified with PyQuant, Proteome Discoverer, and MaxQuant. We found that Proteome Discover and MaxQuant routinely underestimated the heavy label's abundance, resulting in a systematic bias in the SILAC ratios reported. We hypothesize this may be due to the reliance of these algorithms on matching corresponding peaks from each label's isotopic distribution thereby ignoring peaks that deviate significantly from the theoretical distribution of the isotopic cluster (**Fig. 4B**). Although restricting the data to known theoretical distributions is useful for removing contaminating peaks, in cases such as this, it is an incorrect assumption. Thus, to correctly quantify labels that have undergone catabolism, parameters in PyQuant can be set to not enforce the theoretical distribution on labeled data. With this option, PyQuant correctly identifies and quantifies the entire isotopic envelope (**Fig. 4C**). Thus, PyQuant can be particularly useful in correcting

complications arising from metabolic pathways and permits labeling of organisms with minimal loss of quantitative information.

Pseudo label-free quantification of isobaric tagging experiments with PyQuant

In ratio based labeling strategies a common pitfall is the absence of a reference channel that precludes the ability to calculate a ratio for a given PSM. For large, multiplex experiments this is particularly problematic as a missing value in the reference channel prohibits the quantification in many samples. To circumvent this, PyQuant can provide ms^1 and ms^2 based measurements of TMT and iTRAQ experiments. This permits the traditional, ratio-based strategy to be applied and further allows the user to transform each channel into a label-free intensity. We analyzed an experiment that combined known concentrations of iTRAQ labeled A549 cell lysates with this hybrid strategy (Hultin-Rosenberg et al., 2013). We considered how often a chosen reference channel was zero and the corresponding ms^1 based quantification was available. Using this method, we found that depending on which channel was chosen as a reference, the number of quantified PSMs increased by 437 to 965, with an average of 700 PSMs (**Fig. 5**). Thus, PyQuant offers the ability to fill in missing values with a psuedo label-free approach, or to convert a multiplexed iTRAQ or TMT experiment to a label-free like analysis.

Targeted ion quantification with PyQuant

In conventional data-dependent acquisition experiments, the most abundant ions are chosen for fragmentation. While this approach is useful in most cases, it can lead to non-intuitive results in complex samples. For instance, if many ions species exist in a

given scan window that are of similar intensities, the mass spectrometer will randomly select a set number of ions to fragment. In replicate studies, this has the effect that ions selected in one sample may not be chosen for fragmentation in the other sample. This is commonly observed in the sometimes poor overlap in PSMs between replicates. To overcome this issue, PyQuant can perform targeted searches for peptides identified in one replicate that are not fragmented, but present in the ms^1 spectra of another replicate.

As a positive control, we evaluated the ability of PyQuant to correctly predict missing values between two experimental SILAC experimental replicates. One replicate was run through the traditional PyQuant pipeline to establish the true SILAC ratios for each peptide and then the search results of the second replicate were used to quantify peptides in the first replicate. The traditional run provides the true value of a given peptide, and quantifying the first replicate using the search results of the second replicate provides an "interpolated" value for a given peptide. However, because we know the true value for each peptide, the ability of the missing value search to match the first replicate indicates how accurate this method is. This analysis revealed a pearson's correlation of 0.77 between the true values and the values assigned via the "missing value" search (**Supplementary Figure 6**).

To determine how well this method increased the concordance between replicates, we applied PyQuant's missing value analysis to a phosphotyrosine enrichment experiment. A recent study evaluated the changes in phosphorylation levels as a function of TSLP signaling (Zhong et al., 2012). Upon re-analysis of this data, we found between two replicates, one replicate contained 186 phosphotyrosine peptides unique to it and the

second replicate contained 131 phosphotyrosine peptides unique to it (**Fig. 6A**). We performed a targeted search for these missing peptides on the raw data of each replicate with PyQuant. This targeted search provided replicate information for 77 of these phosphopeptides, which increased the concordance quantified phosphopeptides by from 46% to 59% between the replicates (**Fig. 6B**). Lastly, we plotted the phosphopeptide fold changes for each replicate with and without the missing value analysis. This revealed that many peptides followed the same general trend of phosphopeptide abundance between the two samples (**Fig. 6C**).

It is important to note that this approach does have limitations, as many ions can appear at identical *m/z* values and retention times. However, in well controlled experiments, we feel this option in conjunction with PyQuant's detailed visual outputs is a powerful tool for adding confidence to measured peptides which are considered significant in one replicate, but missing in another replicate.

Discussion

As the rate of data acquisition increases in mass spectrometry, GUI based applications that require human intervention, such as MaxQuant and Proteome Discoverer, are inherently limited in their ability to scale with the pace of MS data. Thus, we sought to create a tool that is compatible with existing frameworks as well as able to enable the creation of new high throughput mass spectrometry pipelines. To support the existing frameworks, PyQuant can be provided with the output from MaxQuant, Proteome Discoverer, and the Trans-Proteomic Pipeline. However, to use not explicitly

encoded formats, PyQuant can also be provided with a tab delimited file indicating ions of interest.

We envision PyQuant as a useful complement to existing tools that can be used to provide an additional measure of confidence for quantified values as well as quantify ions that were omitted by a given program. Additionally, by providing a single quantification node that can plug into a variety of data sources and provides a simple, easily parsed output, PyQuant allows the creation of minimal, light weight data analysis pipelines. Moreover, by decoupling the quantification of mass spectrometry data from the traditional peptide spectrum identification and quantification pipeline, PyQuant opens new avenues for novel and highly customized quantification strategies.

Figure Legends

Figure 1: Overview of PyQuant's data processing algorithm.

Figure 2: PyQuant correctly quantifies various quantitative mass spectrometry data methods. A) Two known mixtures of SILAC isotopes in *E. coli* is quantified with three different platforms: PyQuant (red line), MaxQuant (black line), and Proteome Discover (blue line). The expected $\log_2$ ratios are -2, and 2, which are indicated by a dashed green line. All three programs are centered on approximately the same ratios, and have similar peak widths. **B)** PyQuant is able to correctly quantify ^{15}N labeled data. The raw ms^1 spectrum of a peptide from a sample comprised of both ^{15}N and ^{14}N labeled peptides. The ratio of the two species as calculated by PyQuant is shown above, which agrees with the manual interpretation of the relative abundance levels. **C)** A multiplexed experiment that is labeled with NeuCode is shown. The box-plots indicate the distribution of peptides quantified by PyQuant, and the dashed lines indicate known mixture ratios. As shown, the median value, represented by a red line within the box-plot, for each ratio measured is approximately equal to the known mixing ratio. Thus, PyQuant is capable of reproducing the known mixtures of isotopes in the experiment. **D)** PyQuant is used to quantify a TMT 10-plex experiment using ms^3 fragmentation for quantification. On the x-axis are values obtained by *Murphy et al.* in their study, and on the y-axis are the same values obtained by PyQuant. As shown, the values are nearly indistinguishable, providing evidence that PyQuant is capable of ms^3-based quantification.

Figure 3: PyQuant is capable of increasing the quantification rates of MaxQuant and Proteome Discover. A) The output from MaxQuant was analyzed with PyQuant, which resulted in a 44% increase in the number of spectra quantified as compared to MaxQuant alone. Similarly, the output from Proteome Discoverer was analyzed with PyQuant, which resulted in a 68% increase in the number of spectra quantified as compared to Proteome Discover alone. The difference in total spectra between MaxQuant and Proteome Discover is due to each program using different search engines for spectral assignment. **B)** The quantification by PyQuant of a peptide which was not quantified by MaxQuant is shown. The extracted ion chromatogram for each label is robust with multiple isotopomers identified for every label. The quantified ratio as compared to the expected ratio for each channel is shown on the right.

Figure 4: PyQuant quantifies isotopic profiles resulting from amino acid conversions. A *C. elegans* sample was labeled with Arginine-10, which was metabolically converted to other amino acids. As seen, this results in a spread of the heavy isotopic cluster as compared to the light isotopic cluster. This results in Proteome Discover quantifying only a subset of the heavy data, and systematically under-reporting Heavy/Light SILAC ratios. PyQuant is capturing and quantifying most of the spread isotopic cluster.

Figure 5: Pseudo label-free quantification with PyQuant. PyQuant is used to quantify both ms^2 and ms^1 data of an iTRAQ experiment. PyQuant provides an additional ~400-1000 quantified PSMs in the dataset, contingent upon which channel was chosen as a reference.

Figure 6: Filling in missing values between replicate studies. A) The Venn diagram shows the overlap of hyperphosphorylated peptides from two replicate phosphotyrosine enrichments. Over half of the data is not replicated between the two samples. **B)** After performing a missing value search with PyQuant on peptides unique to each replicate, peptides which were not fragmented in a given replicate are quantified. This increase is shown by the increased overlap between the two replicates. **C)** The SILAC ratios of phosphopeptides between each replicate. Each axis represents the SILAC ratios of a particular replicate. Blue circles correspond to phosphopeptides fragmented and quantified in both replicates (the overlap in **A**). Red circles represent the values recovered from PyQuant's missing value analysis. Each red circle was fragmented in only one replicate, but by using the elution time and precursor ion, the ion is able to be quantified irrespective of its fragmentation. As shown, the red and blue circles have a similar trend between the two replicates, indicating the missing values follow a similar trend as the ions fragmented in both replicates.

Supplementary Figure 1: Re-calibration of mass error as a function of *m/z*. A) The mass error of identified peptides is plotted as a function of the peptide's precursor ion. As shown, there is an increase in the mass error as a function of the precursor mass. In green is a spline function used to fit this data. **B)** Using the spline function, the masses from panel A are re-calibrated to remove the systematic bias of mass error as a function of *m/z*.

Supplementary Figure 2: A Gaussian mixture model is used to isolate interfering ions. Shown is raw mass spectrometry data in profile mode. The precursor mass of interest is at 800.9035, but there is a co-eluting ion which was also detected. To remove

this, a Gaussian mixture model is applied to the data, and the true signal of the ion of interest can be extracted, which is shown by the red line.

Supplementary Figure 3: Various types of extracted ion chromatograms. Examples of the profiles of various extracted ion chromatograms are shown. In orange is the raw intensity values of a given ion species as a function of retention time and in blue is the curve PyQuant fits to the data. The black line is the retention time of the corresponding ms^2 scan for each ion. This illustrates the difficulties encountered when fitting a function to the elution profile of an ion, and how PyQuant is capable of accurately fitting an assortment of data.

Supplementary Figure 4: PyQuant identifies and removes outlier peaks. For each panel, the raw and fitted extracted ion chromatograms of a peptide is shown. In orange is the raw intensity of each isotopic peak as a function of retention time, with the fitted curve shown in blue. Each panel represents the isotopic peaks of an identified peptide. Because the elution profile of each isotopic peak should be similar, PyQuant evaluates how similar each fitted peak is. The isotopic peak in panel **A** could be better fit by a more skewed peak; however, because this skewness is not present in the isotopic peaks shown in panel **B** and **C**, the final fit is a less skewed peak.

Supplementary Figure 5: PyQuant's confidence assessment selects for more accurate data. Three confidence thresholds for a dataset with a known mixing ratio is shown. The data should be centered at 1, and the width of each box-plot shows the deviation of the quantified peptides from this value. As peptides below a given

confidence threshold indicated on the x-axis are excluded, the deviation of the peptides from the true value decreases.

Supplementary Figure 6: Correlated of true values versus interpolated values from PyQuant's missing value analysis. On the x-axis are the true Heavy/Light peptide ratios from a replicate. The y-axis are these peptides quantified using the retention time and precursor ion mass of a replicate. There is a high concordance between the true values and interpolated values for each peptide, indicating the missing value analysis is accurate.

Table 1: Mixing ratios of SILAC isotopes

Figure 1

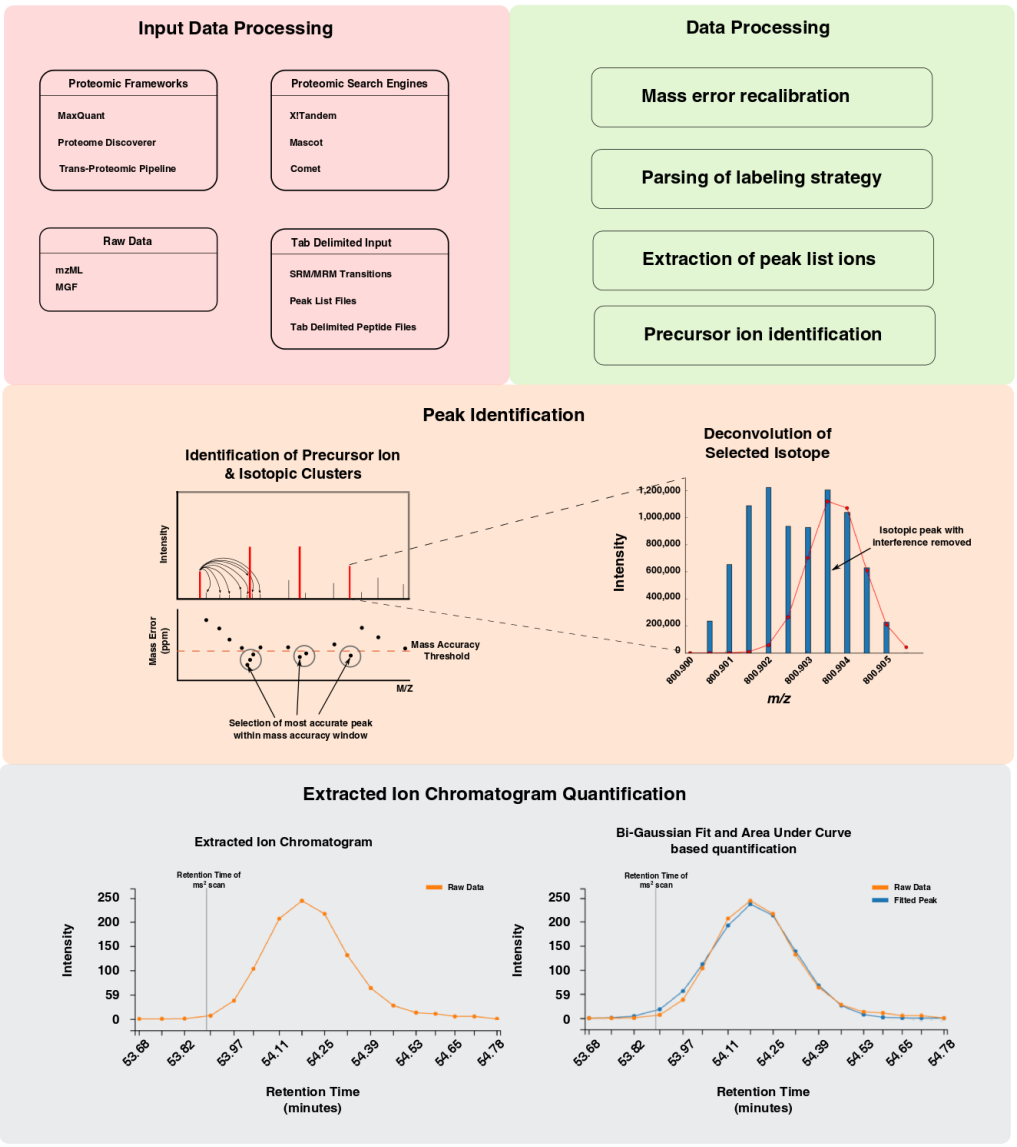

Input Data Processing
Proteomic Frameworks
MaxQuant
Proteome Discoverer
Trans-Proteomic Pipeline
Proteomic Search Engines
X!Tandem
Mascot
Comet
Raw Data
mzML
MGF
Tab Delimited Input
SRM/MRM Transitions
Peak List Files
Tab Delimited Peptide Files
Data Processing
Mass error recalibration
Parsing of labeling strategy
Extraction of peak list ions
Precursor ion identification
Peak Identification
Identification of Precursor Ion & Isotopic Clusters
Intensity
Mass Error (ppm)
Mass Accuracy Threshold
M/Z
Selection of most accurate peak within mass accuracy window
Deconvolution of Selected Isotope
Intensity
Isotopic peak with interference removed
m/z
Extracted Ion Chromatogram Quantification
Extracted Ion Chromatogram
Retention Time of ms² scan
Raw Data
Intensity
Retention Time (minutes)
Bi-Gaussian Fit and Area Under Curve based quantification
Raw Data
Fitted Peak

Figure 2

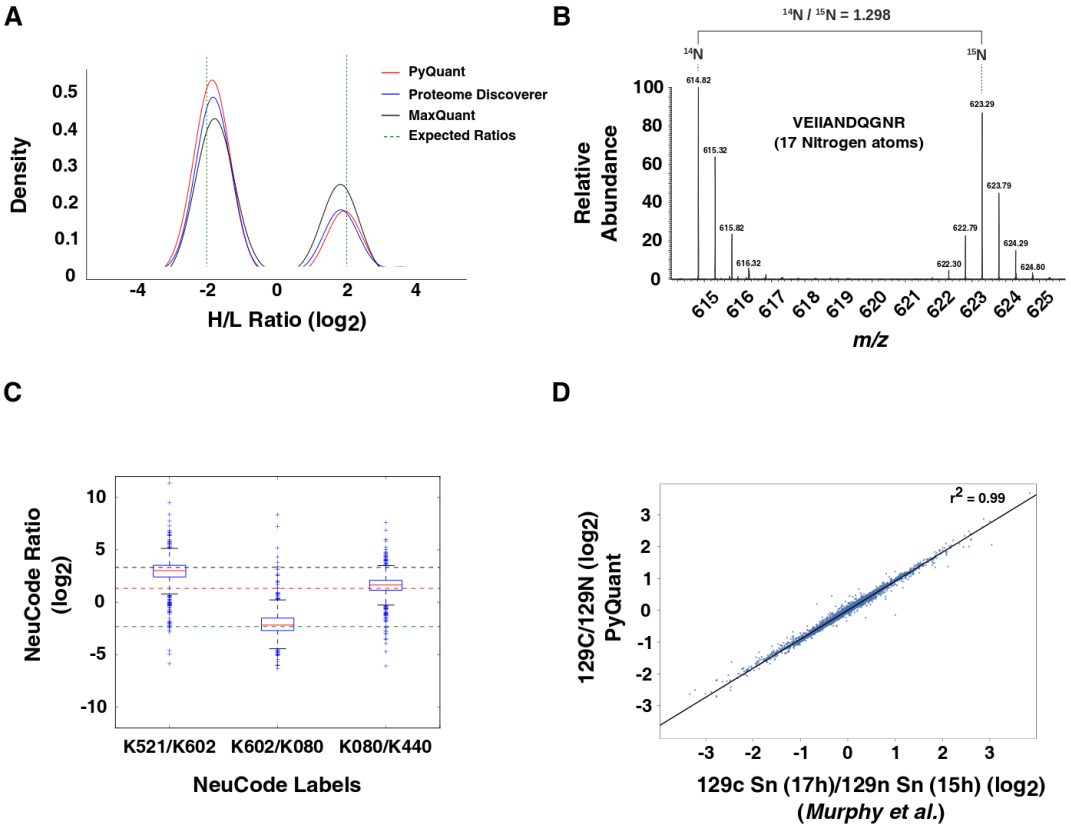

A
Density
H/L Ratio (log2)
PyQuant
Proteome Discoverer
MaxQuant
Expected Ratios
B
14N / 15N = 1.298
VEIIANDQGNR
(17 Nitrogen atoms)
Relative Abundance
m/z
C
NeuCode Ratio (log2)
K521/K602
K602/K080
K080/K440
NeuCode Labels
D
r2 = 0.99
129C/129N (log2) PyQuant
129c Sn (17h)/129n Sn (15h) (log2)
(Murphy et al.)

Figure 3

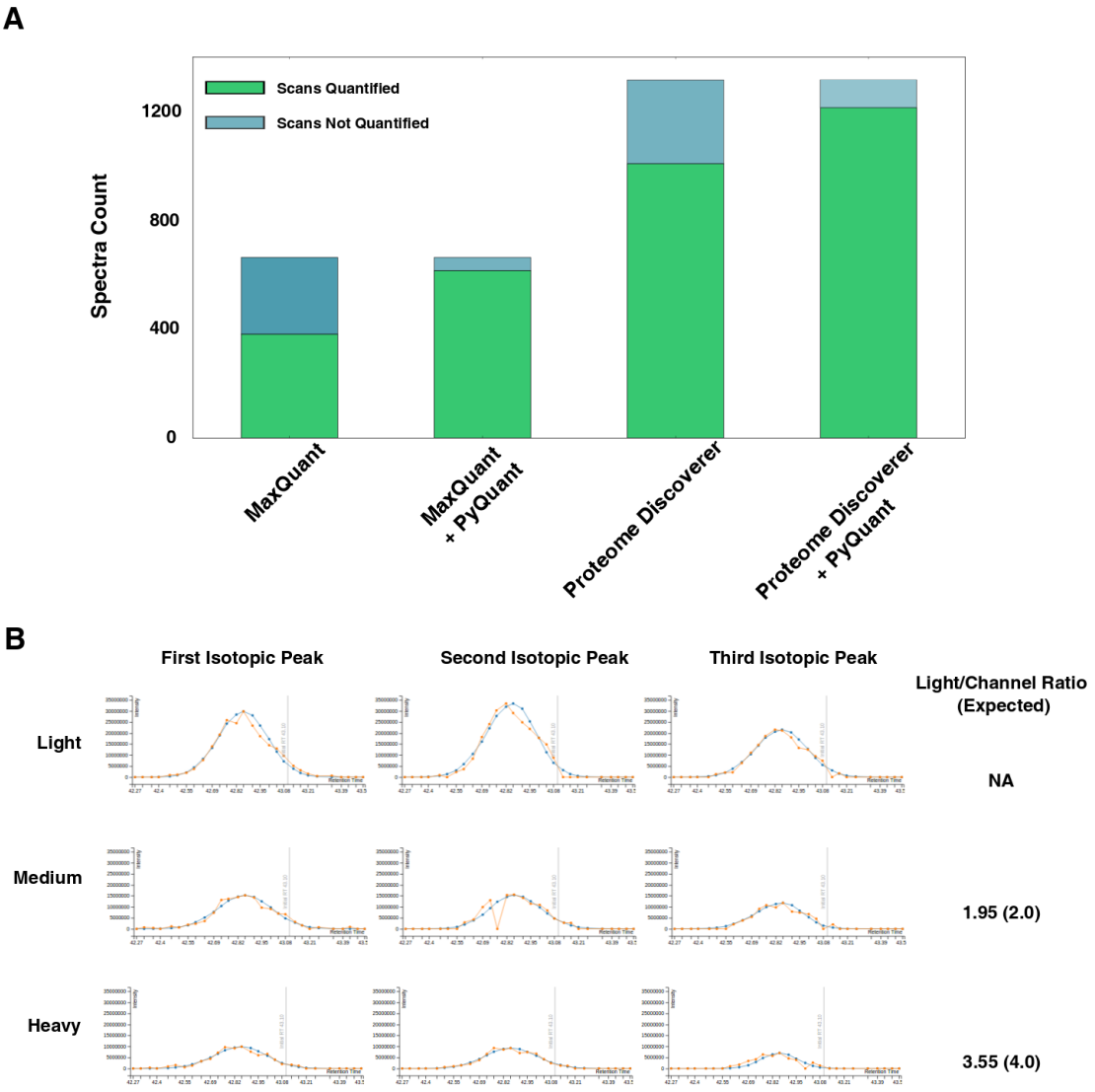

A
Scans Quantified
Scans Not Quantified
Spectra Count
1200
800
400
0
MaxQuant
MaxQuant + PyQuant
Proteome Discoverer
Proteome Discoverer + PyQuant
B
First Isotopic Peak
Second Isotopic Peak
Third Isotopic Peak
Light/Channel Ratio (Expected)
Light
NA
Medium
1.95 (2.0)
Heavy
3.55 (4.0)

Figure 4

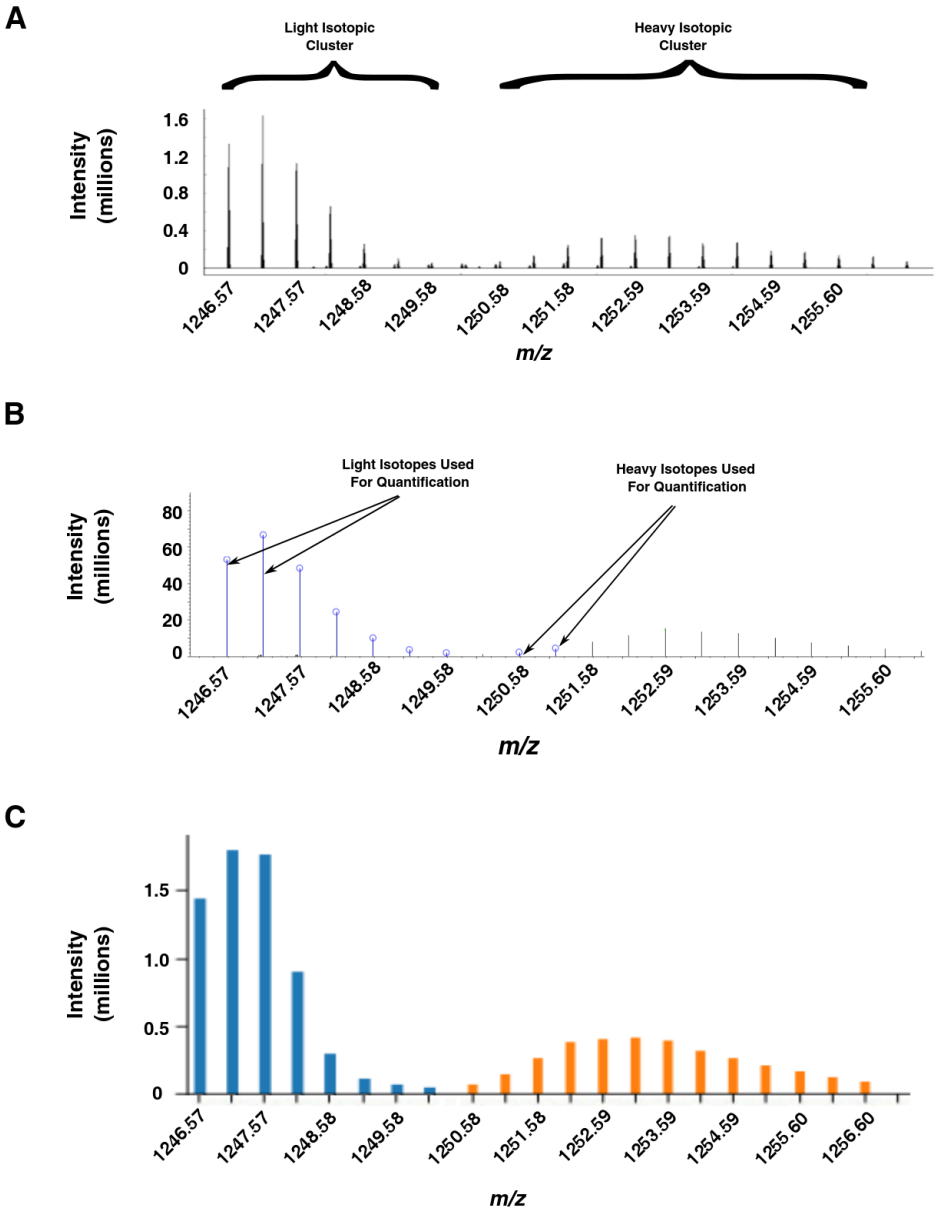

A
Light Isotopic Cluster
Heavy Isotopic Cluster
Intensity (millions)
1.6
1.2
0.8
0.4
0
1246.57
1247.57
1248.58
1249.58
1250.58
1251.58
1252.59
1253.59
1254.59
1255.60
m/z
B
Light Isotopes Used For Quantification
Heavy Isotopes Used For Quantification
Intensity (millions)
80
60
40
20
0
1246.57
1247.57
1248.58
1249.58
1250.58
1251.58
1252.59
1253.59
1254.59
1255.60
m/z
C
Intensity (millions)
1.5
1.0
0.5
0
1246.57
1247.57
1248.58
1249.58
1250.58
1251.58
1252.59
1253.59
1254.59
1255.60
1256.60
m/z

Figure 5

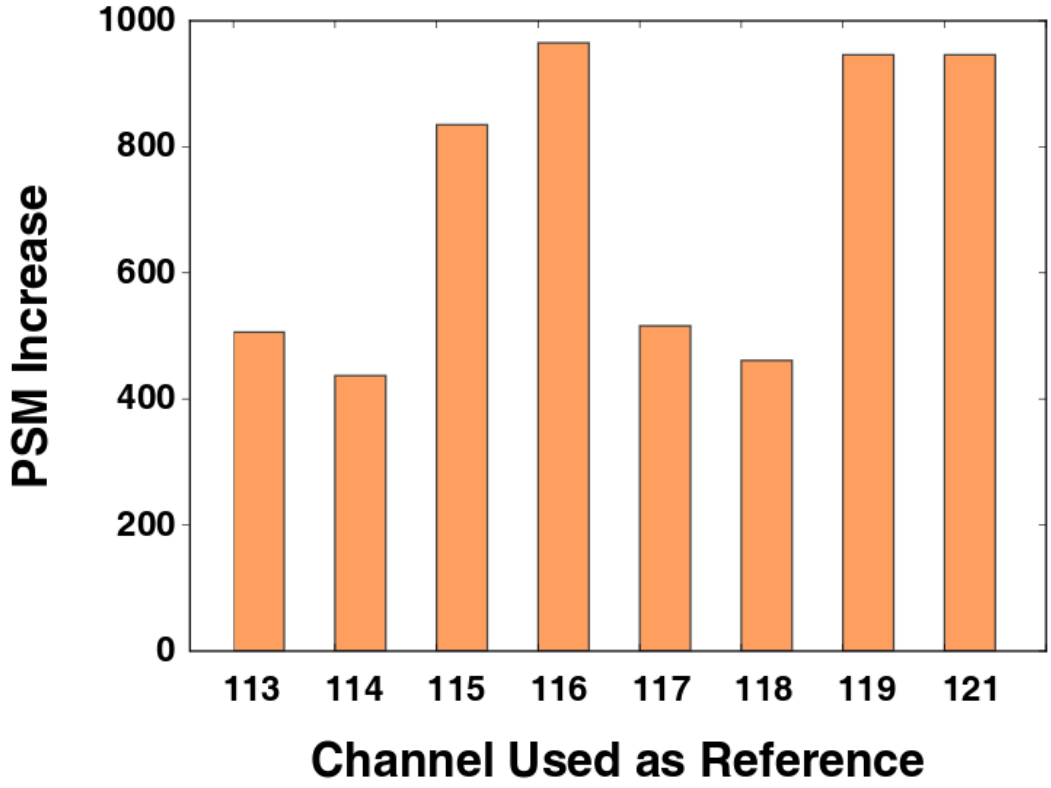
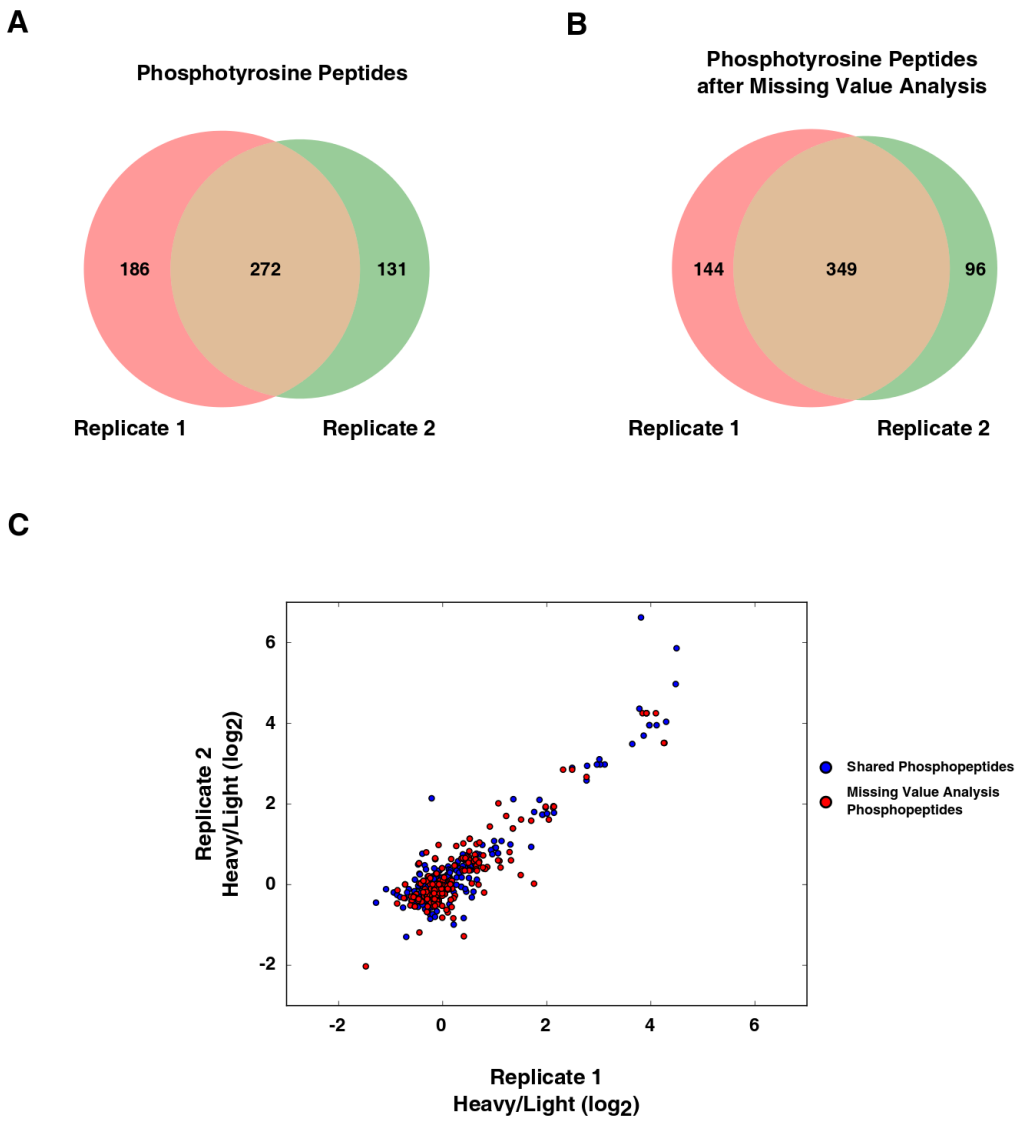
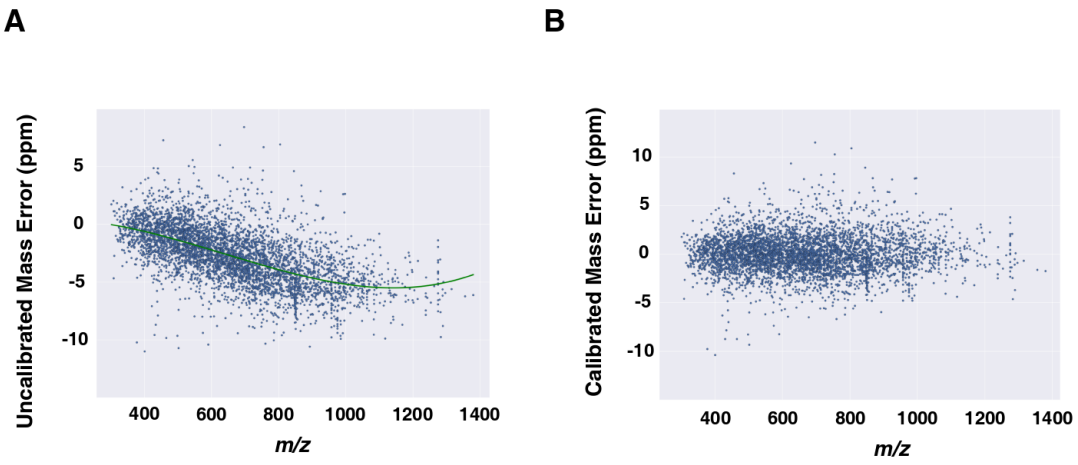

PSM Increase
1000
800
600
400
200
0
113
114
115
116
117
118
119
121
Channel Used as Reference

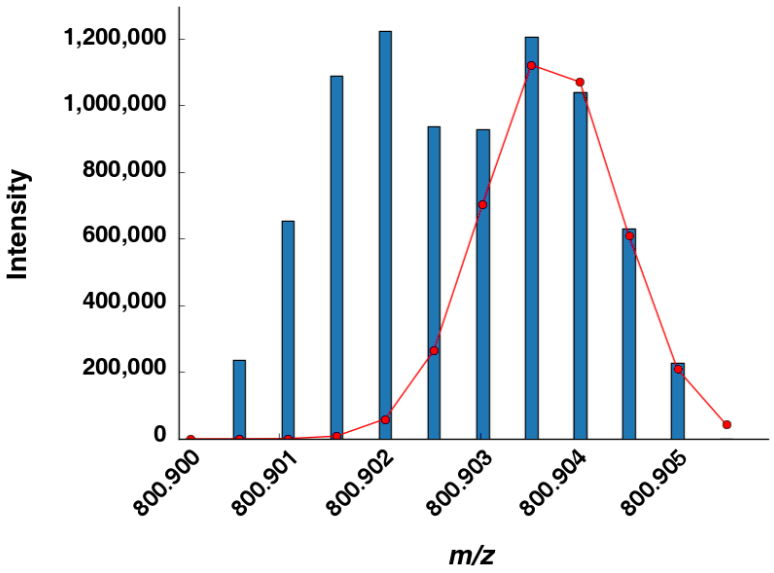
Figure 6


A
Phosphotyrosine Peptides
186
272
131
Replicate 1
Replicate 2
B
Phosphotyrosine Peptides
after Missing Value Analysis
144
349
96
Replicate 1
Replicate 2
C
Replicate 2
Heavy/Light (log2)
Replicate 1
Heavy/Light (log2)
Shared Phosphopeptides
Missing Value Analysis
Phosphopeptides

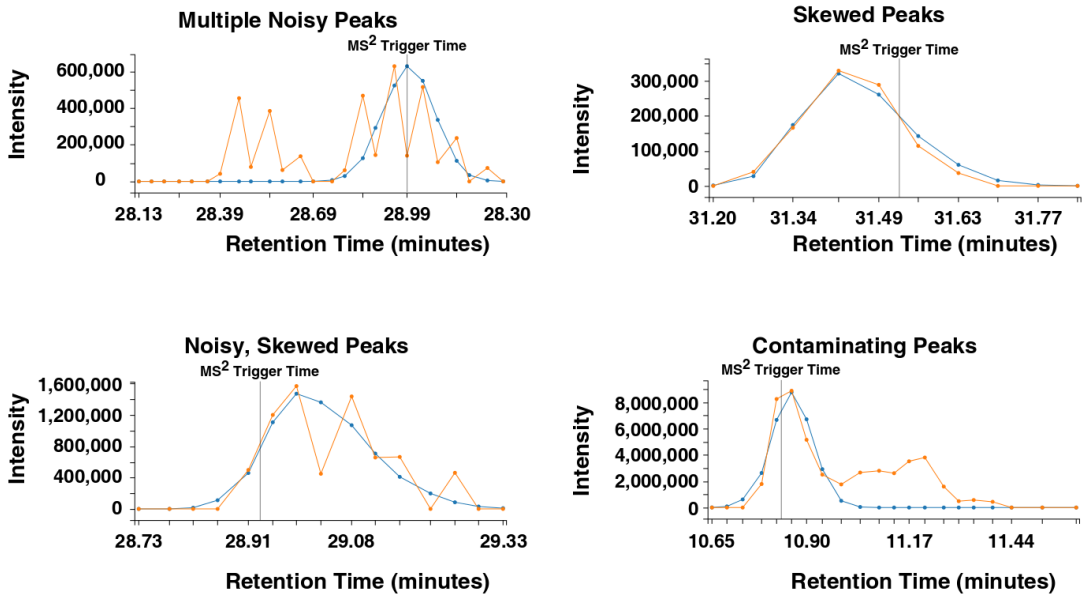
Supplementary Figure 1



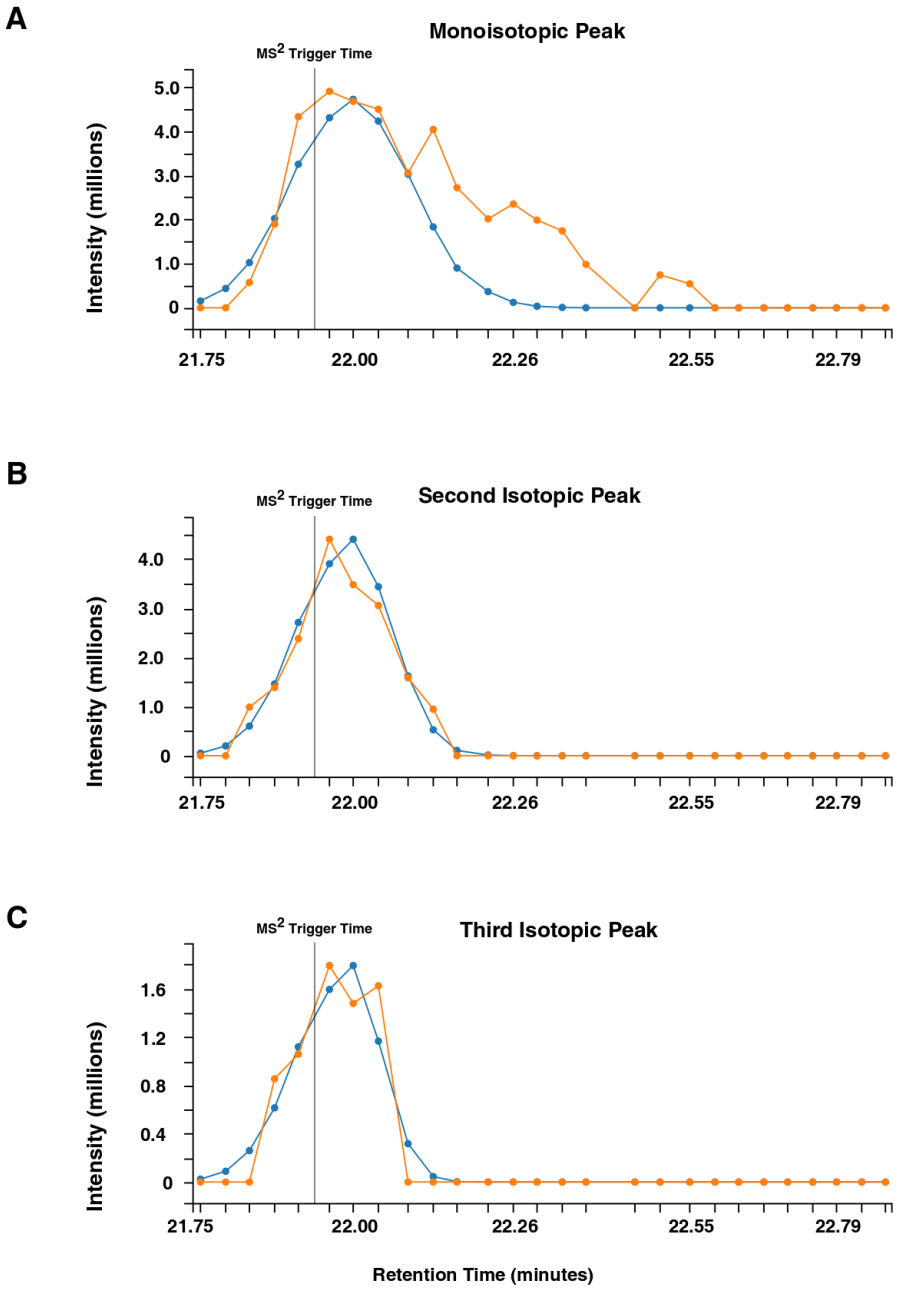
Supplementary Figure 2



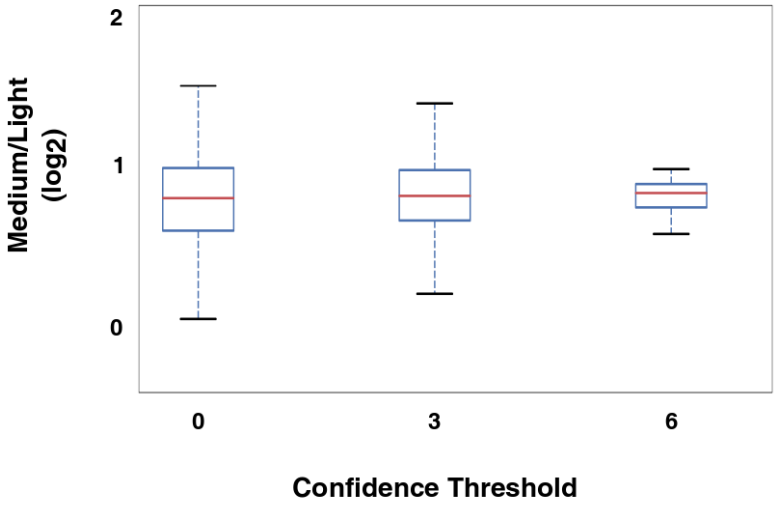
Supplementary Figure 3



Supplementary Figure 4



Supplementary Figure 5



Supplementary Figure 6

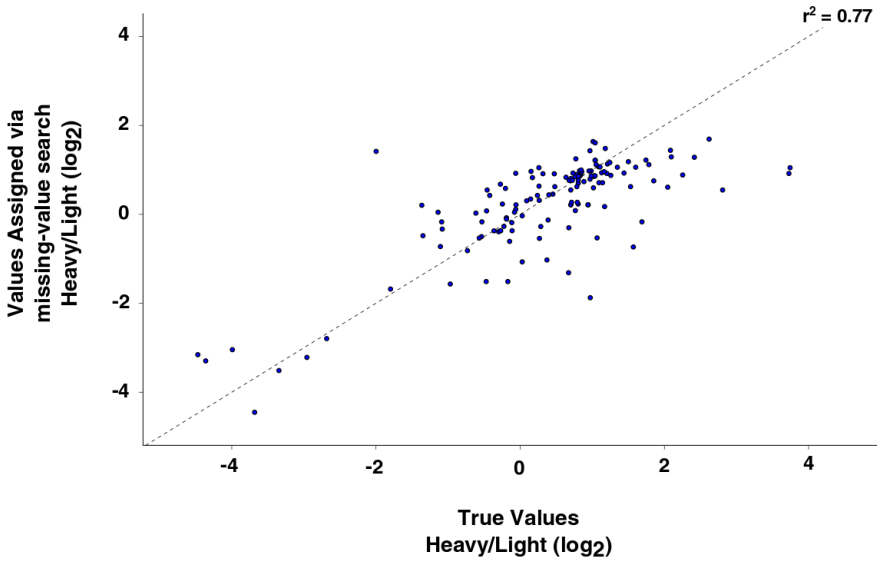


Table 1: Samples grown with known ratios of isotopically labeled amino acids

	Mixing Ratios					
	Lys-0	Lys-4	Lys-8	Arg-0	Arg-6	Arg-10
Sample 1	4	0	1	1	0	4
Sample 2	1	0	1	1	0	1
Sample 3	1	2	4	4	2	1

CHAPTER 3

Metabolic conversion of isotopically labeled amino acids

Although SILAC has predominantly been utilized in cell cultures, it has been increasingly introduced into many model organisms for global, quantitative analysis of protein abundance. In the previous chapter, I briefly touched upon a problem PyQuant solved: the ability to handle isotopic conversions of labeled amino acids. The conversion of isotopically labeled amino acids by the sample has been a known potential complication of SILAC (Ong et al., 2003; Van Hoof et al., 2007). In normal cell culture experiments it has empirically been found to not be a significant issue; however, in organisms it has been observed and the conversion of the isotope has a significant impact on the data (Borek et al., 2015; Larance et al., 2011). To further explain this concept, and why it matters – what an isotopic cluster is what exactly comprises the cluster will be explained in further detail here.

First, we must consider the elemental composition of a peptide sequence. A peptide sequence is primarily composed of four elements: carbon, oxygen, nitrogen, and hydrogen. An element has a natural distribution of isotopes that are present in varying proportions, with the lightest isotope usually being the dominant species. When an amino acid is being synthesized, there is no bias in which isotope is incorporated and accordingly, there is a random mixture of the isotopes for each element with the likelihood of each isotope present being proportional to the natural abundance of that isotope. For mass spectrometry, let's imagine there are no isotopes and we are analyzing

the peptide sequence PEPTIDE, which contains 34 carbons, 54 hydrogens, 7 nitrogens, and 15 oxygens. If on a mass spectrometer, we measure this pure singly charged species, we would observed a single peak at a *m/z* of 800.36724 (**Figure 1A**).

Now, if we include isotopes, additional peaks corresponding to different combinations of isotopes appear. If we have a single heavy isotope in the peptide, a new peak at a *m/z* value corresponding to an additional neutron will be added. An important technical detail is that because of mass defects, the additional neutron is not the same mass for every element. However, for simplicity, in our treatment we will assume there is no mass defect and all neutrons are of equal mass. In our example, an isotope of any element can be incorporated into an amino acid at any position. Similarly, there is a probability for amino acids being incorporated that contain multiple isotopes. The net effect of this is for a peptide to spread from its monoisotopic mass to a distribution that reflects the probability of the peptide sequence containing one, two, three, and more isotopes (**Figure 1B**). One important feature is as the peptide sequence becomes longer, the monoisotopic peak becomes increasingly smaller, and with long sequences may be indistinguishable from noise.

In traditional SILAC experiments, both heavy and light peptide sequences have an equal abundance of naturally occurring isotopes in their media. Thus, the isotopic distribution of the heavy and light species should be similar (**Figure 2A**). Without amino acid conversions, this pattern can be theoretically calculated and used to identify which peaks correspond to a peptide of interest. Additionally, this information can be used to increase the accuracy of selecting the isotopes corresponding to a peptide of interest. If a

peak deviates significantly from the expected value, it is likely to be a contaminating ion. However, when there is an additional source of labeled isotopes, such as metabolic conversion of labeled amino acids, the isotopic cluster widens and deviates significantly from the theoretical distribution, thus complicating the interpretation of protein abundance (**Figure 2B**).

To mitigate this conversion, several approaches have been employed. One approach is the use of genetic manipulation to prevent the conversion of arginine. In *C. elegans*, the ornithine transferase gene was inhibited by RNAi (Larance et al., 2011) and deleted in yeast (Bicho et al., 2010) to disrupt this metabolic pathway. Another approach has been to limit the isotopes used for SILAC labeling to lysine, which is not known to be metabolically converted. This has been applied in several organisms including *C. elegans (Fredens et al., 2011), M. musculus* (Zanivan et al., 2012), and *D. melanogastar* (Sury et al., 2010). Though these methods are useful for avoiding metabolic conversions, they all possess concessions and are situational. Knocking down or deleting genes involved in metabolic pathways will have side effects, and creation of a knockdown or knockout in a given organism may not be feasible. Utilizing only lysine limits digestion to Lys-C, which is significantly more expensive than trypsin, and leaves many peptides longer than is optimal for mass spectrometry.

Although experimental methods to avoid metabolic conversions are useful, they are stopgap solutions that mask the fundamental issue – that existing computational tools are unable to handle metabolic conversions. Many existing tools rely on the theoretical isotopic peak pattern to correctly identify the isotopic cluster and provide accurate

SILAC ratios. When these assumptions are invalid, many peptides are incorrectly quantified or not quantified at all due to the exclusion of the full isotopic envelope of the labeled peptide (**Figure 2B**). PyQuant was created to surpass this limitation and accommodate metabolic conversion of labeled amino acids for accurate SILAC quantification. Thus, PyQuant allows for SILAC in organisms without the use of genetic manipulation or introducing additional variables such as RNAi. Through the use of PyQuant, the amount of peptides quantified in *C. elegans* was drastically increased. On a single dataset, Proteome Discoverer was capable of quantifying ~13,000 peptides while PyQuant quantified over 60,000 peptides and when processing the output of MaxQuant, MaxQuant quantified ~21,000 peptides while PyQuant quantified over 40,000 peptides (**Figure 3**). The ability to quantify peptides with this unexpected isotopic distribution enabled the quantification of our *C. elegans* data and allows us to pursue our original goal – the identification of phosphatase substrates in an organism.

Figure Legends

Figure 1: An isotopic distribution of a peptide with and without isotopes. A) A peptide that is composed of a single isotopic species for all elements contained with the peptide. **B)** A peptide where the natural isotopic abundance of ^{13}C and ^{15}N is taken into account. The peptide's abundance is spread over several distinct peaks which correspond to the peptide sequence with one or more isotopes present. Notably, the monoisotopic peak has diminished in size.

Figure 2: Comparison of a light and heavy SILAC ratio with and without metabolic conversion. A) A "normal" SILAC pair is shown that was derived from a sample where metabolic conversion of the isotopically labeled heavy amino acids was not occurring. The filled blue dots indicate the isotopic peaks that were used to estimate the ion abundance of the light and heavy labeled peptides. **B)** A SILAC pair where the heavy label was isotopically converted and incorporated into other amino acids. This results in the spread of the isotopic label, as can be seen by comparing the width of the heavy cluster as compared to the width of the light cluster. Because of this deviation, the program used to quantify the data, Proteome Discoverer, only included the first two isotopic peaks from the heavy label.

Figure 3: PyQuant significantly increases the number of quantified peptides in the presence of amino acid conversions. A comparison of the numbers of peptides quantified of a dataset where conversion of the heavy isotope occurred. The output of

MaxQuant and Proteome Discoverer are processed with PyQuant, which increases the number of quantified peptides.

Figure 1

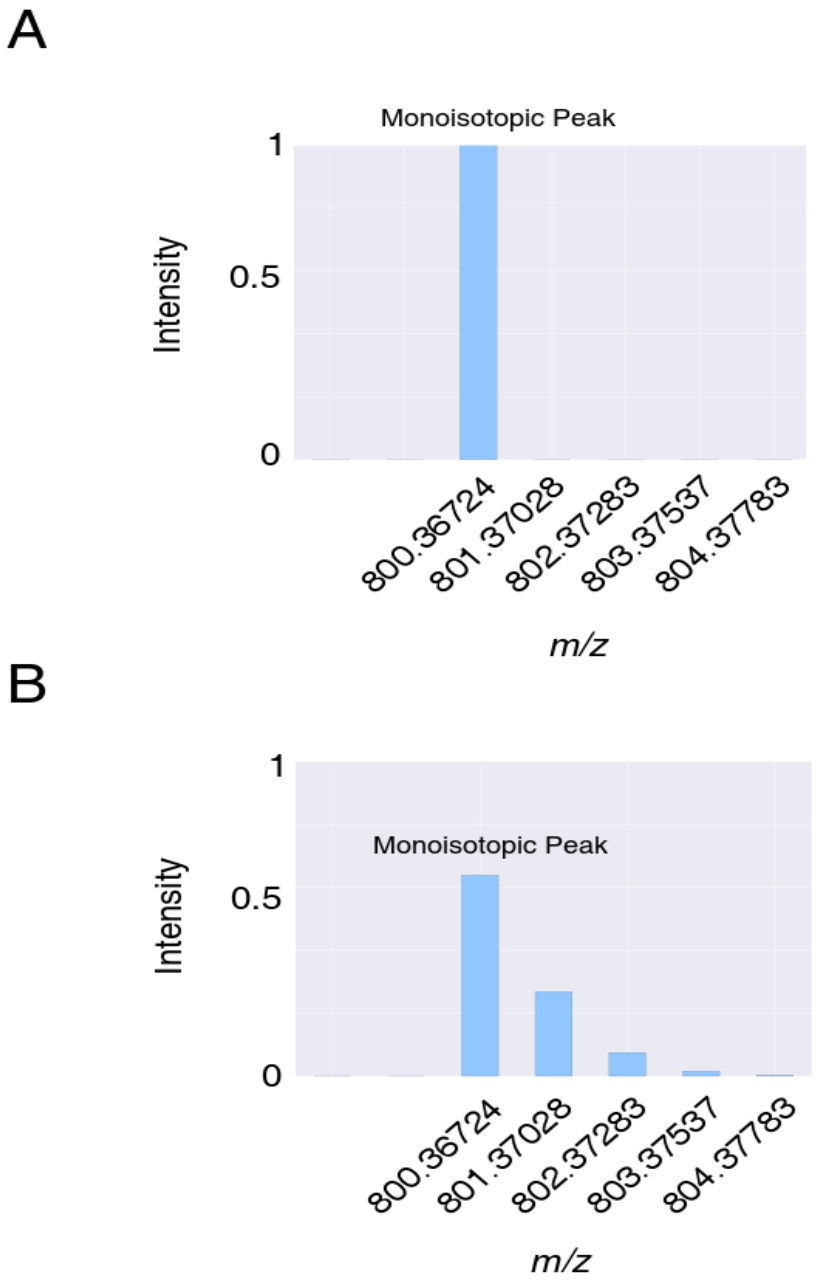

A
Monoisotopic Peak
1
0.5
0
Intensity
800.36724
801.37028
802.37283
803.37537
804.37783
m/z
B
1
0.5
0
Intensity
Monoisotopic Peak
800.36724
801.37028
802.37283
803.37537
804.37783
m/z

Figure 2

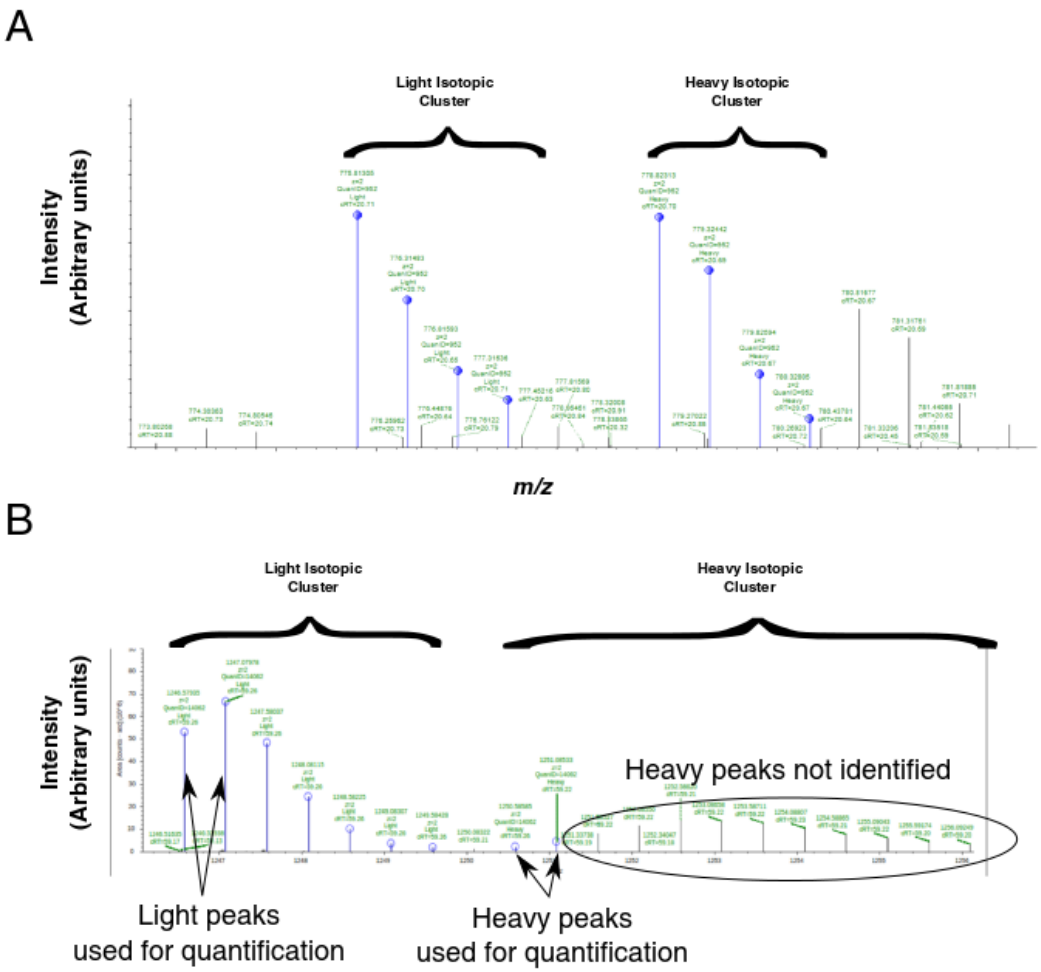
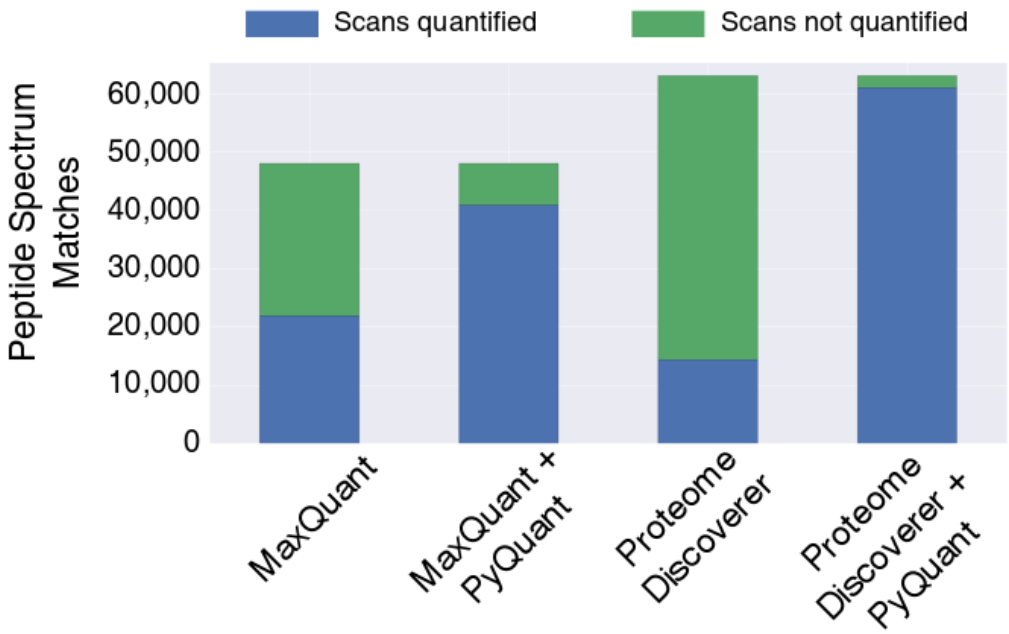

A
Light Isotopic Cluster
Heavy Isotopic Cluster
Intensity (Arbitrary units)
m/z
B
Light Isotopic Cluster
Heavy Isotopic Cluster
Intensity (Arbitrary units)
Heavy peaks not identified
Light peaks used for quantification
Heavy peaks used for quantification

Figure 3


Scans quantified
Scans not quantified
Peptide Spectrum Matches
60,000
50,000
40,000
30,000
20,000
10,000
0
MaxQuant
MaxQuant + PyQuant
Proteome Discoverer
Proteome Discoverer + PyQuant

CHAPTER 4

Unbiased identification of substrates of protein tyrosine phosphatase ptp-3 in C. elegans

Introduction

Tyrosine phosphorylation is a reversible post-translational modification involved in most cellular processes. Dysregulation of tyrosine phosphorylation has been implicated in a number of human cancers (Hanahan and Weinberg, 2000). The role of tyrosine kinases as potent oncogenes has been well described. For example, *HER2* is found to be amplified in 20% of breast cancers (Slamon et al., 1987), the BCR-ABL translocation is observed in nearly all chronic myelogeneous leukemias (Buchdunger et al., 2001) and increased expression of *EGFR* is reported in 40-80% of non-small cell lung cancers and 50% of primary lung cancers (Kolibaba and Druker, 1997). Protein tyrosine phosphatases, on the other hand, have been shown to act as tumor suppressors. Functional inactivation of protein tyrosine phosphatases through either mutation or decreased expression is associated with a multitude of cancers (Jacob and Motiwala, 2005; Li et al., 1997; The Cancer Genome Atlas Network, 2012; Veeriah et al., 2009). Recently, protein tyrosine phosphatases have also been found to play a role as oncogenes in addition to acting as tumor suppressors (Tonks, 2006). For example, deletion of the protein tyrosine phosphatase non-receptor type 11 (*PTPN11*) gene results in increased development of hepatocellular carcinoma in mice (Bard-Chapeau et al., 2011) whereas activating

mutations or overexpression has been described in several forms of leukemia (Chan et al., 2008).

Though both protein tyrosine kinases and protein tyrosine phosphatases play integral roles in cancer biology, there is a marked discrepancy in our knowledge regarding phosphatases as compared to kinases. For instance, although there are approximately equal numbers of tyrosine phosphatases and tyrosine kinases in the human genome, the phosphorylation database PhosphoSitePlus lists over 1,700 known tyrosine kinase substrates while the phosphatase database DEPOD contains ~130 tyrosine phosphatase substrates (Duan et al., 2015; Hornbeck et al., 2015). This disparity in our knowledge is partly due to intrinsic differences in the experimental methodologies that are amenable to studying kinases versus phosphatases. Kinases result in a gain of signal that more easily lend themselves to experimental techniques such as radiolabeling ATP to identify substrates, or more recently, using mass spectrometry to identify phosphorylated sites. Phosphatases, on the other hand, lead to a loss of signal and cannot be studied in a straightforward fashion. One approach to identifying phosphatase substrates is the creation of substrate trapping mutants in which a catalytically inactive phosphatase is used to pull-down interacting proteins and methods such as western blotting and mass spectrometry are used to identify substrates. While useful, this approach is an *in vitro* binding method and requires a functional, mutated fusion protein which may not be possible to generate easily.

To expand upon the known phosphatase substrates, we sought to apply a high-throughput strategy that would lead to discovery of novel substrates of tyrosine

phosphatases *in vivo*. We reasoned that functionally inactive mutants or knockdowns of a phosphatase would result in an increase in the phosphorylation levels of its substrates. This, combined with mass spectrometry, could rapidly identify hyperphosphorylated peptides derived from putative substrates of a protein phosphatase. Moreover, because mass spectrometry is used to detect phosphopeptides, the actual sites that are differentially phosphorylated due to a mutation are revealed.

For this study, we chose to identify substrates of the LAR family of protein tyrosine phosphatases. These phosphatases are receptor-like tyrosine phosphatases characterized by several extracellular fibronectin and immunoglobulin domains along with two intracellular tyrosine phosphatase domains. Within vertebrates, there are at least three related genes comprising the LAR family: the LAR receptor protein tyrosine phosphatase (*PTPRF*), protein tyrosine phosphatase receptor type delta (*PTPRD*), and protein tyrosine phosphatase receptor type sigma (*PTPRS*). *PTPRD* has been shown to be a tumor suppressor gene and has been found to be functionally inactivated in glioblastoma multiforme, head and neck cancer, and lung cancer (Veeriah et al., 2009). Loss of *PTPRS*, which is known to dephosphorylate *EGFR*, has been reported in ~25% of head and neck cancers and its deletion in lung cancer results in increased resistance to the tyrosine kinase inhibitor erlotinib (Morris et al., 2011).

Given the association between mutations in the LAR family and a number of cancers, identification of substrates of the LAR family may help elucidate the molecular consequences of these mutations and how they might lead to oncogenesis. However, such studies in mouse have been complicated by the functional redundancy between the

multiple members of the LAR family. This redundancy can be mitigated through the use of multiple gene knockouts; however, existing attempts to knockout multiple members of the LAR family have resulted in embryonic lethality (Uetani et al., 2006). To circumvent this issue, we used *C. elegans* as a model system to identify phosphatase substrates in an organism which has reduced redundancy but retains many important developmental effects of the loss of LAR. The single LAR family homolog in *C. elegans, ptp-3,* is made up of the characteristic extracellular fibronectic and immunoglobulin domains and two intracellular tyrosine phosphatase domains that are observed in this family. Similar to its human homologs, *ptp-3* has been implicated in gastrulation, epidermal development and embryonic morphogenesis. Further, deletion of *ptp-3* results in smaller brood size due to a diminished embryonic viability (Harrington et al., 2002). Here, through the use of quantitative phosphoproteomics by employing SILAC labeling of *C. elegans*, we identified 255 putative substrates of *ptp-3* and used computational methods to identify common motifs for this phosphatase. Further, we identified that kinases were overrepresented as substrates of *ptp-3,* with their kinase activation loop being a common target of dephosphorylation by *ptp-3*.

Materials and Methods

Reagents

The *E. coli SLE1* strain and all *C. elegans* strains were purchased from the *Caenorhabditis* Genetics Center (CGC). Heavy labeled amino acids - arginine ($^{13}C_6$-$^{15}N_4$) and lysine ($^{13}C_6$-$^{15}N_2$) were purchased from Cambridge Isotopes Laboratories (Andover,

MA). For phosphotyrosine enrichment, anti-phosphotyrosine mouse monoclonal antibody (P-Tyr-1000) was purchased from Cell Signaling Technology (Danvers, MA). Nystatin (N1638-100ML) was purchased from Sigma-Aldrich. Complete, EDTA-free protease inhibitor cocktail tablets (10711400) were purchased from Roche. Sequencing grade trypsin and lys-c enzymes were purchased from Promega and TPCK-treated trypsin was purchased from Worthington Biochemical Corp. All other reagents were purchased from Sigma.

C. elegans metabolic labeling

For SILAC labeling of *C. elegans*, the bacterial food source was labeled and then fed to the *C. elegans*. For culturing bacteria, light cultures were grown in Lysogeny Broth (LB) and heavy cultures were grown in M9 Minimal Media supplemented with 40 µg/ml final concentrations of heavy arginine and heavy lysine. Each culture was grown overnight, and centrifuged at 8,000 x *g* for 10 minutes at 4ºC. For each 250 ml of culture pelleted, 50 ml of S. Medium (Stiernagle, 2006) was added to resuspend the pellet. *C. elegans* were added to a 1 L flask containing 50 ml of resuspended *E. coli* supplemented with 10 µg/ml of Nystatin to prevent fungal growth and grown according to previously described methods (Stiernagle, 2006). The culture was monitored and when the bacteria was almost clear (approximately 3 days), the *C. elegans* were pelleted and resuspended in 0.1 M NaCl. To complete purification, an equal volume of 60% sucrose was added and centrifuged for 5 minutes at 900 x *g* at 4ºC. *C. elegans* at the surface were then pipetted off, and washed three times in 50 ml of 0.1 M NaCl to obtain a pure, live population of *C. elegans*.

C. elegans lysis and protein digestion

C. elegans were resuspended in 9M Urea containing a protease inhibitor cocktail (Roche 10711400), 50 mM beta-galactosidase, 25 mM NaF, and 1 mM Sodium Orthovanadate. Next, the resuspension was sonicated and viewed under a microscope following sonication until the *C. elegans* were sufficiently broken apart. The lysate was cleared by centrifugation at 14,000 x *g* for 20 minutes at 4°C. For whole proteome experiments, 250 µg from wild type and *ptp-3* lysates were combined, reduced with 5 mM of dithiothreitol, and alkylated with 5 mM of iodoacetamide. Then, the lysate was diluted to a final urea concentration of 3M and digested with lys-c with an enzyme to protein ratio of 1:100 for 4 hours at room temperature. After digestion with lys-c, the crude digests were diluted to a final urea concentration of 1.5M and digested overnight with sequencing grade trypsin with an enzyme to protein ratio of 1:25. After confirming digestion efficiency, the sample was acidified with 1% trifluoroacetic acid (TFA) and centrifuged at 2000 x *g* for 5 minutes at room temperature. The supernatant was loaded onto a Sep-pak C_{18} cartridge (Waters, Cat#WAT051910) and equilibrated with 0.1% TFA. Desalted peptides were then lyophilized and subjected to basic reverse phase liquid chromatography. For phosphotyrosine peptide enrichment, 25 mg each from heavy and light lysates were combined, reduced with 5 mM of dithiothreitol, and alkylated with 5 mM of iodoacetamide. Then, lysates were diluted to a final urea concentration of 1.5 M and digested overnight with TPCK-treated trypsin (Worthington) with an enzyme to protein ratio of 1:25. Peptide digests were desalted with a Sep-Pak C cartridge (Waters,

18

Cat#WAT051910) and then lyophilized for subsequent use in phosphotyrosine enrichment.

Immunoaffinity Purification of Tyrosine Phosphopeptides

Immunoaffinity purification (IAP) of tyrosine phosphopeptides was carried out as previously described (Kim et al., 2014b; Rush et al., 2005). Briefly, after lyophilization, 50 mg of peptide mixture was dissolved in 1.4 ml of IAP buffer (50 mM MOPS pH 7.2, 10 mM sodium phosphate, 50 mM NaCl) and subjected to centrifugation at 2000 × *g* at room temperature for 5 min. Before IAP, P-Tyr-1000 beads (Cell Signaling Technology) were washed with IAP buffer twice at 4 °C and the pH of the supernatant containing peptides was adjusted to 7.2 by adding 1 M Tris Base. For IAP, the supernatant was incubated with P-Tyr-1000 beads (Cell Signaling Technology) at 4°C for 30 min, and the beads were washed three times with IAP buffer followed by two washes with pure water. Enriched peptides were eluted twice from beads by incubating the beads with 0.15% TFA at room temperature.

LC-MS/MS

Peptide samples were analyzed on an LTQ-Orbitrap Elite mass spectrometer (Thermo Electron, Bremen, Germany) interfaced with Easy-nLC II nanoflow liquid chromatography systems (Thermo Scientific) as previously described (Kim et al., 2014a). Peptides were reconstituted in solvent A (0.1% formic acid) and loaded onto a trap column (75 µm × 2 cm) packed in-house with Magic C18 AQ (Michrom Bioresources, Inc., Auburn, CA, USA) (5 µm particle size, pore size 100 Å) at 280 bar with solvent A

(0.1% formic acid in water). Peptides were resolved on an analytical column (75 μm × 15 cm) at a flow rate of 300 nl/min using a linear gradient of 5–35% solvent B (0.1% formic acid in 95% acetonitrile). Mass spectrometry analysis was carried out in a data dependent manner with full scans (300–1,700*m/z*) acquired using an Orbitrap mass analyzer at a mass resolution of 120,000 at 400 *m/z*. The ten and fifteen most intense precursor ions from a survey scan were selected for MS/MS from each duty cycle for whole proteome and phosphopeptides, respectively, and detected at a mass resolution of 30,000. All the tandem mass spectra were produced using higher energy C-trap dissociation method. Dynamic exclusion was set for 30 s with a 7 p.p.m. mass window. Internal calibration was carried out using lock-mass from ambient air (*m/z* 445.1200025).

Data Analysis

Tandem mass spectra were processed and searched using MASCOT (Version 2.2.0) (Perkins et al., 1999) and SEQUEST (Eng et al., 1994) search algorithms against a *C. elegans* WormBase (Harris et al., 2010) database (version 248) through the Proteome Discoverer platform (version 2.0, Thermo Scientific). For both algorithms, the search parameters included a maximum of two missed cleavages; carbamidomethylation at cysteine as a fixed modification; N-terminal acetylation at protein N-termini, deamidation at asparagine and glutamine, oxidation at methionine, phosphorylation at serine, threonine and tyrosine and SILAC labeling at $^{13}C_6{}^{15}N_2$ lysine and $^{13}C_6{}^{15}N_4$ arginine as variable modifications. The mass tolerance for peptide and fragment ions were set to 10 ppm and 0.1 Da, respectively. The false discovery rate was set to 1% at the peptide level.

The probability of phosphorylation for each Ser/Thr/Tyr site on each peptide was calculated by the PhosphoRS node in Proteome Discoverer 2.0.

Following the database search, SILAC ratios were quantified with PyQuant and a 2-fold cutoff was chosen to indicate a significant change in a given PSM between samples. For whole proteome measurements, the data was normalized to the median to control for loading bias. After normalization, peptide spectra matches (PSMs) were grouped at the peptide level and the median value was taken of each peptide group. Following this, the median value of a protein's peptide groups was taken to estimate the baseline protein abundance. Next, in house python scripts were used to provide protein level assignments of phosphotyrosine sites (e.g. VLPAATY(phos)LK becomes CE000001 Tyr582). To normalize loading bias during mixing of samples, the data were normalized against each enrichment's corresponding whole proteome experiment. Identified phosphotyrosine sites were then grouped together, and the median value of these groups was used to measure the phosphorylation level of each phosphotyrosine site. Finally, the abundance level of each site was normalized against the protein abundance measured from the whole proteome experiment. If possible, proteins were then mapped to human homologs and the corresponding human residues of phosphotyrosine sites were annotated.

For mapping *C. elegans* genes to human homologs, the Wormbase API was used. To perform Gene Ontology based clustering and functional annotation, the R package clusterProfiler (Yu et al., 2012) and the PANTHER classification system (Mi et al., 2013) were used, respectively. For identification of motifs, the surrounding sequence (the 13

flanking residues) from each phosphotyrosine site was extracted from the WormBase protein database and analyzed using MEME (Bailey and Elkan, 1994) and GLAM2 (Frith et al., 2008) to generate motifs. For prediction of sequences containing these motifs, motifs generated with MEME were searched using FIMO and predictions with a q-value below 10^{-4} were retained (Grant et al., 2011). For motifs generated using GLAM2, GLAM2SCAN was used to predict sequences containing the motif and predictions with a score above 15 were kept (Frith et al., 2008).

Data Availability

Mass spectrometry derived proteomics data has been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository with the dataset identifier PXD003233.

Results

Deletion of the ptp-3 phosphatase results in increased global tyrosine phosphorylation

A mutant *C. elegans* strain harboring a mutation in *ptp-3,* designated RB1878*,* was previously generated by the *C. elegans* Gene Knockout Consortium. This mutant contains a premature stop codon leading to the truncation of *ptp-3* prior to its phosphatase domains. For quantitative measurements, wild type worms were labeled with heavy isotope-labeled lysine and arginine amino acids, and mutant worms were labeled with light versions of these amino acids. Following a label check to confirm incorporation of the heavy isotope, samples were lysed and protein concentration was measured. From each sample, 25 mg of lysate was taken, mixed, and digested in-solution with trypsin and

subjected to an antibody based phosphotyrosine enrichment and LC/MS-MS analysis (**Figure 1**) as described previously by our group (Syed et al., 2015; Zahari et al., 2015; Zhong et al., 2012). Peptides were identified using Mascot and Sequest run through Proteome Discoverer 2.0 and quantified using PyQuant.

Two biological replicates of wild type and *ptp-3* mutants each were grown, lysed, and subjected to phosphotyrosine enrichment and whole proteome analysis. First, we assessed proteomic changes of the *ptp-3* knockout on *C. elegans*. From our analysis, we identified 2,130 proteins encoded by 1,266 genes, with 148 proteins encoded by 72 genes that were upregulated and 29 proteins encoded by 20 genes that were downregulated in the *ptp-3* knockout (**Supplementary Figure 1A**). To assess the possible impact of the knockout, we performed a Gene Ontology (GO) analysis (Ashburner et al., 2000) using the PANTHER classification system (Mi et al., 2013). GO analysis of genes revealed an enrichment for genes involved with developmental processes (**Supplementary Figure 1B**), which is consistent with the known phenotype of increased embryonic lethality and abnormal morphology in the *ptp-3* knockout.

Using antibody-based phosphotyrosine peptide enrichment, we identified 3,021 phosphotyrosine containing peptides. We grouped the identified peptides by their phosphotyrosine site (e.g. y543). This analysis identified 1,463 different phosphotyrosine sites. Because global protein changes in *ptp-3* knockout worms may be falsely inferred as changes in phosphorylation levels, protein levels determined from whole proteome analysis were used to correct for protein fluctuations that may contribute to phosphotyrosine measurements. We considered a site hyper or hypophosphorylated if

there was at least a 2-fold change in the phosphorylation state between the wild type and mutant samples. Using these criteria, we identified 255 sites that were hyperphosphorylated and 264 that were hypophosphorylated in the *ptp-3* mutant (**Figure 2**). As a phosphatase, knockout of *ptp-3* predicts an increase in phosphorylation levels of its substrate(s); therefore, we consider these 255 hyperphosphorylated sites to be potential substrates of *ptp-3*.

To evaluate the general impacts of the loss of *ptp-3* mediated dephophorylation, we analyzed hyperphosphorylated tyrosine sites in the *ptp-3* knockout. A GO analysis of putative substrates was performed using clusterProfiler (Yu et al., 2012). Similar to the whole proteome, GO analysis revealed enrichment of genes related to reproduction, anatomical development and embryo development.

Conservation of putative substrates in humans

Due to evolutionary changes from *C. elegans* to *H. sapiens*, putative substrates identified in *C. elegans* may not be conserved across species. Moreover, conservation is a strong predictor of functionality, and sites that are conserved across species are more likely to be true positives than those that are found only in *C. elegans* (Tan et al., 2009). Thus, for our subsequent analysis we focused only on putative substrates that were conserved in *C. elegans* and humans.

We performed a search using Wormbase (Harris et al., 2010) to locate homologous genes of the identified putative *ptp-3* substrates. This analysis revealed 145 putative substrates with clearly defined human homologs. Next, Clustal Omega (Sievers

et al., 2011) was used to align each *C. elegans* protein sequence to its human homolog and to determine if the specific phosphotyrosine site was conserved in humans. Of the 145 putative substrates with homologs in humans, 46 contained a tyrosine residue in both species at the identified hyperphosphorylated tyrosine site. Interestingly, we also found that in 11 cases the tyrosine residue that was hyperphosphorylated in *C.* elegans was replaced by a phenylalanine in the orthologous sequence in humans. This is of note because mutational studies often mutate tyrosine residues to phenylalanine in order to prevent phosphorylation yet retain the structural importance of aromatic side chains. Thus, these residues may not be crucial for signaling in humans, but could have important roles in structure-based interactions.

Site level resolution of known substrates of the LAR family

To evaluate the success of our profiling approach, we compared the list of proteins that were hyperphosphorylated in the *ptp-3* mutant with *C. elegans* orthologs of previously described substrates of the LAR family. We chose to use the substrates of the LAR family members in the human dephosphorylation database DEPOD (Duan et al., 2015) because currently there are no databases listing dephosphorylation events in C. elegans, there is no entry for *ptp-3* in the Worm Interactome Database (Simonis et al., 2009), and we are unable to find known substrates of *ptp-3* in the literature through PubMed searches.

At this time, the DEPOD database contains nine known substrates of the LAR family, with three substrates belonging to the same gene family. Of the known substrates,

only Tyr705 of *STAT3* was annotated as the site that is dephosphorylated by *PTPRD*. Remarkably, in our data, we identified orthologs of four of the nine known substrates, *NTRK2***,** *MET***,** *INSR***,** and *STAT3* (sites indicated in **figure 2**), to be hyperphosphorylated with a conserved tyrosine residue. Additionally, the *C. elegans* ortholog of *STAT3, sta-1*, was found and the *C. elegans* site corresponding to Tyr705, Tyr588, exhibited a 5.6 fold increase in phosphorylation levels in the *ptp-3* knockout. This shows we are able to find known gene products which are dephosphorylated by the LAR family in addition to the site which has previously been shown to be dephosphorylated.

The DEPOD database listed only one substrate where the actual tyrosine residue that is dephosphorylated was confirmed. This is an important limitation to known enzyme-substrate relationships; that even when experimental evidence shows a protein is dephosphorylated by a phosphatase, the exact sites which are modified are often not known. Since mass spectrometry was used to identify substrates in this study, it inherently provides site level resolution as well as quantitative measures of phosphorylation. As discussed above, in our data three substrates, *NTRK2*, *MET*, and *INSR*, were known to be dephosposphorylated in humans although their actual site or sites of dephosphorylation was unknown. We identified conserved hyperphosphorylated sites mapping to these three genes as well as a hyperphosphorylated site that has been previous described. This illustrates that our approach is able to provide highly confident sites for future functional studies that further dissect phosphorylation based signaling networks.

Putative substrates of ptp-3 are enriched in kinases

Next, we performed a GO analysis for biological processes of the conserved putative substrates of *ptp-3*, which revealed a significant enrichment for kinases (q-value $5.9 * 10^{-27}$). Of these genes included in the GO analysis, we found that 15 conserved hyperphosphorylated sites were located in kinases. We then examined the location of the phosphosite in each kinase, and found that 6 sites were in the activation loop and have previously been shown to be critical for kinase activation (**Figure 3**). Additionally, several sites were identified that were not in the activation loop, but conferred important catalytic activity. For example, Tyr579 of *PDGFRB* and Tyr15 *CDK5* were hyperphosphorylated, and phosphorylation of these sites has been shown to increase kinase activity (Baxter et al., 1998; Zukerberg et al., 2000). Tyr1356 of *MET* was hyperphosphorylated, which is not known to increase kinase activity, but has been shown to be crucial for transduction of signals involving cell motility, morphogenesis, and tubule formation (Dekel et al., 2003).

Next, we sought to determine whether a common motif exists within the putative substrates of *ptp-3. W*e selected the identified conserved hyperphosphorylated tyrosine sites and 13 amino acids surrounding each site as the input sequences for motif prediction. These sequences were scanned for common motifs using GLAM2 (Frith et al., 2008) and MEME (Bailey and Elkan, 1994), two motif identification programs that are part of the MEME suite (Bailey et al., 2009). This analysis revealed two motifs where the hyperphosphorylated tyrosine residue was located at the 0 position (**Figure 4A; 4B**). Next, we used these motifs to predict additional substrates of *ptp-3* and the LAR family by searching against the human RefSeq protein database with GLAM2SCAN (Frith et al.,

2008) and FIMO (Grant et al., 2011), two programs that take the respective output of GLAM2 and MEME and identify sequences containing a given motif. The motif identified by GLAM2 (**Figure 4A)**, was present in 65 proteins encoded by 20 genes while the MEME motif (**Figure 4B)** was identified in 108 proteins encoded by 29 genes. Significantly, 12 of the 20 genes containing the GLAM2 motif (60%) and all 29 genes containing the MEME motif are kinases, implying an important role for *ptp-3* and the LAR family in the regulation of kinases.

Given the abundance of kinases identified through our profiling approach, and that many hyperphosphorylated tyrosine residues were responsible for increased kinase activity, this suggests that functional inactivation of a LAR family member likely leads to increased kinase activity. Taking into account the known oncogenic potential of kinases, this data presents a possible link between the mutation of various LAR family members and associated cancers (FUNATO et al., 2011; Giefing et al., 2011; Morris et al., 2011; Veeriah et al., 2009). Although our profiling approach was able to recapitulate many known substrates of the human homolog of *ptp-3* and provided broad implications to the regulatory roles of *ptp-3* and the LAR family in humans, it should be noted these are putative substrates and do not supplant directed mutational studies and *in vitro* assays for final confirmation. However, given the ability of this approach to capture much of what is known, we feel the hyperphosphorylated peptides identified in this study are excellent candidates for further validation including functional studies to precisely map out the LAR family's dephosphorylation network.

Discussion

The *C. elegans* LAR-like receptor tyrosine phosphatase, *ptp-3*, is expressed in several tissues during *C. elegans* development and has been shown to play a role in gastrulation and neural morphogenesis (Chin-Sang et al., 2002; Harrington et al., 2002). A human homolog of *ptp-3*, *PTPRD*, has been found to be deleted or functionally inactivated in several cancers, with inactivation of *PTPRD* occurring in over 50% of GBM tumors (Veeriah et al., 2009). A recent study identified Tyr705 of *STAT3* as a specific target of *PTPRD*. In our data, we found the *C. elegans* STAT orthologs in humans to be hyperphosphorylated at Tyr588, which upon alignment to the human *STAT3* protein, corresponds to Tyr705 of *STAT3*. Interestingly, another receptor tyrosine phosphatase, *PTPRK*, has been shown to dephosphorylate Tyr705 of *STAT3* as well (Chen et al., 2015). Given that multiple homologs of *ptp-3* exist in humans, this may indicate that putative substrates found in our study may be common to multiple protein receptor type tyrosine phosphatases.

Many substrates of the LAR-like receptor tyrosine phosphatases are known only at the protein level and lack site level resolution. *PTPRF* has been shown to dephosphorylate the insulin receptor (*INSR*), beta-catenin, and hepatocyte growth factor receptor (*MET*) but there is no site level resolution (Hashimoto et al., 1992; Machide et al., 2006; Müller et al., 1999). In our data, we find hyperphosphorylated tyrosine residues corresponding to the *C. elegans* homologs of *INSR* and *MET*. In the *C. elegans* homolog of *INSR*, *daf-2*, we find two hyperphosphorylated sites, Tyr1419 and Tyr1420, which are both conserved in humans. For *MET*, we find several hyperphosphorylated sites in its *C. elegans* homolog *svh-2* as well. While only one of these sites is conserved in humans, we

note that there are nearby tyrosine residues which may be the functionally equivalent site in humans. *PTPRS* has been shown to dephosphorylate the epidermal growth factor receptor (*EGFR*) and members of the neutrotrophic tyrosine kinase receptors *NTRK1*, *NTRK2*, and *NTRK3* (Faux et al., 2007; Vijayvargia et al., 2004). We find hyperphosphorylated tyrosine residues in both of the *C. elegans* homologs of these genes. For *EGFR*, like *MET*, the hyperphosphorylated residue is not conserved in the human homolog, but once again a conserved tyrosine residue is near the site identified in *C. elegans*. For the neutrotrophic tyrosine kinase receptors, we identified a hyperphosphorylated Tyr632 in the *C. elegans* homolog *trk-1*, which is conserved in the human sequence as well.

When considering the known literature, our study was able to identify known proteins dephosphorylated by the LAR-like receptor tyrosine phosphatases, and provided potential sites of substrates known to be dephosphorylated by *PTPRF* and *PTPRS*. Additionally, we provide an extensive catalog of potential substrates of *ptp-3* in *C. elegans*, many of which are related to the known developmental roles of *ptp-3* in *C. elegans*. In humans, many of the potential substrates were found to be kinase-related, with 9 of the hyperphosphorylated tyrosine residues having known effects on kinase activity, which may explain why functional inactivation of *PTPRD*, *PTPRF*, and *PTPRS* have been found in multiple cancers. Given the numerous knockouts in *C. elegans*, our strategy can be applied to identify potential substrates of additional phosphatases as well as other post-translational modifications to further characterize signaling networks in an organismal setting.

Figure Legends

Figure 1: Overall strategy of phosphotyrosine peptide enrichment and LC-MS/MS analysis. The experimental workflow for labeling of *C. elegans,* phosphotyrosine enrichment and LC-MS/MS is shown. First, *E. coli* cultures were grown with SILAC isotopes and then fed to *C. elegans* over several generations until the worms were fully labeled. Following label incorporation, wild type and mutant samples were lysed and an equal protein amount of each labeled lysate was mixed together. Following mixing, one portion of the sample was subjected to whole proteome LC-MS/MS analysis and the remaining portion was subjected to phosphotyrosine peptide enrichment and LC-MS/MS analysis.

Figure 2: Distribution of phosphotyrosine sites identified. The relative change in phosphorylation levels of phosphotyrosine sites between the *ptp-3* mutant and wild type is shown. Higher values on the y-axis correspond to increases in phosphotyrosine levels with a tyrosine phosphatase mutation. The red dashed line indicates a 2 fold increase in tyrosine phosphorylation, and the blue dashed line indicates a 2 fold decrease in tyrosine phosphorylation. Shapes with their corresponding gene are known substrates of human homologs of *ptp-3,* and show that this study identifies known phosphotyrosine substrates as being hyperphosphorylated.

Figure 3: Activating residues of kinases that are hyperphosphorylated. Tyrosine residues with increased phosphorylation in the *ptp-3* knockout are depicted, with the identified residue indicated in red. The first six rows correspond to tyrosine residues that

are located in the activation loop of the indicated gene, and the last three rows correspond to tyrosine residues that have been shown to alter kinase activity.

Figure 4: Motif analysis of hyperphosphorylated tyrosine sites. A) A motif identified by MEME for hyperphosphorylated tyrosine residues. The zero position for the motif indicates the hyperphosphorylated tyrosine residue. **B)** A motif identified by GLAM2 for hyperphosphorylated tyrosine residues. The zero position for the motif indicates the hyperphosphorylated tyrosine residue.

Supplementary Figure 1: Whole proteome analysis of the *ptp-3* mutant. A) The distribution of protein abundance between the *ptp-3* mutant and wild type strain measured at the global proteome level is shown. **B)** A GO analysis of genes up-regulated in the whole proteome data shows an enrichment for genes involved in developmental processes, in agreement with the known phenotype of increase embryonic lethality of the *ptp-3* mutant.

Figure 1

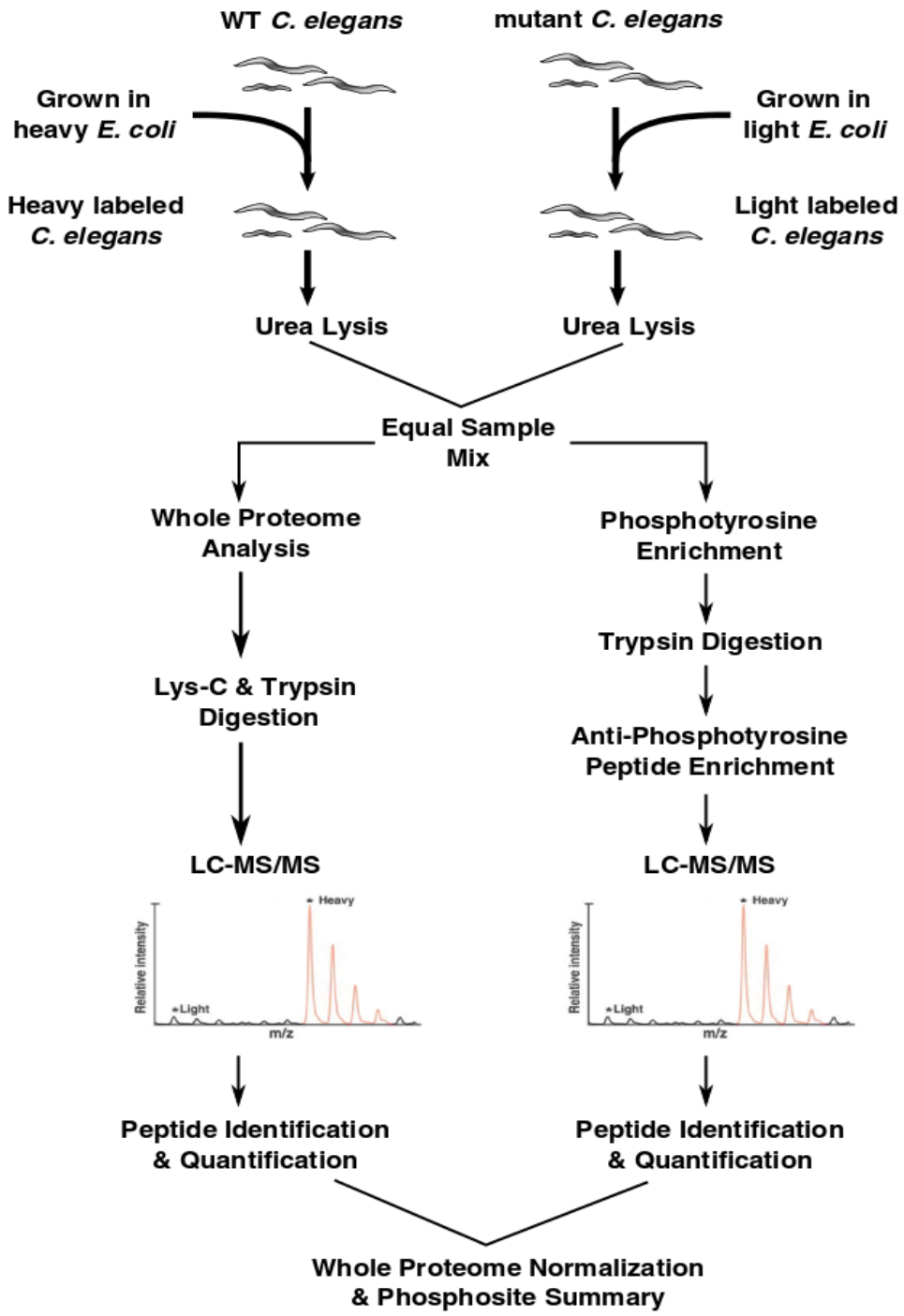

WT C. elegans
mutant C. elegans
Grown in heavy E. coli
Grown in light E. coli
Heavy labeled C. elegans
Light labeled C. elegans
Urea Lysis
Urea Lysis
Equal Sample Mix
Whole Proteome Analysis
Phosphotyrosine Enrichment
Trypsin Digestion
Lys-C & Trypsin Digestion
Anti-Phosphotyrosine Peptide Enrichment
LC-MS/MS
LC-MS/MS
Heavy
Light
Relative intensity
m/z
Heavy
Light
Relative intensity
m/z
Peptide Identification & Quantification
Peptide Identification & Quantification
Whole Proteome Normalization & Phosphosite Summary

Figure 2

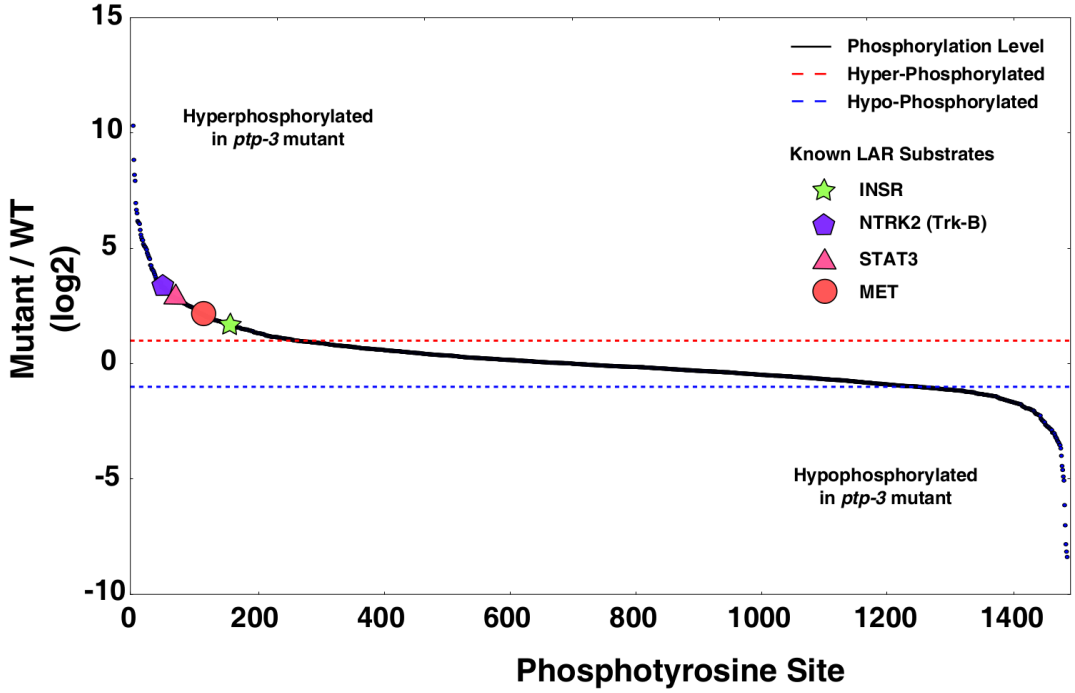

Phosphorylation Level
Hyper-Phosphorylated
Hypo-Phosphorylated
Known LAR Substrates
INSR
NTRK2 (Trk-B)
STAT3
MET
Hyperphosphorylated
in ptp-3 mutant
Hypophosphorylated
in ptp-3 mutant
Mutant / WT
(log2)
Phosphotyrosine Site
15
10
5
0
-5
-10
0
200
400
600
800
1000
1200
1400

Figure 3

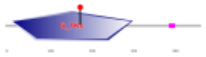
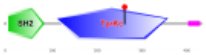
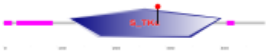
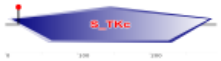
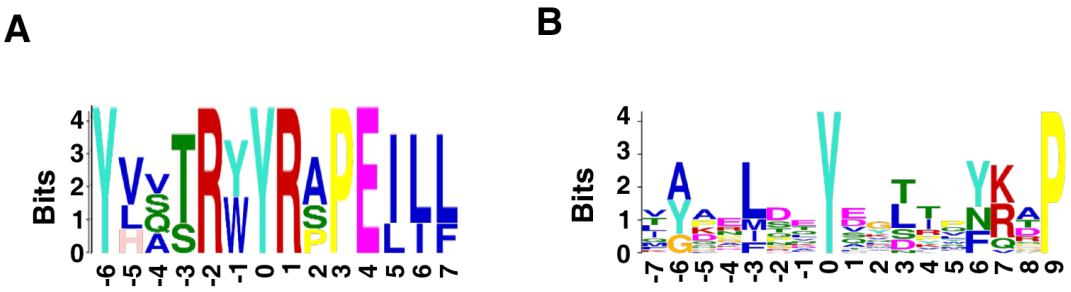
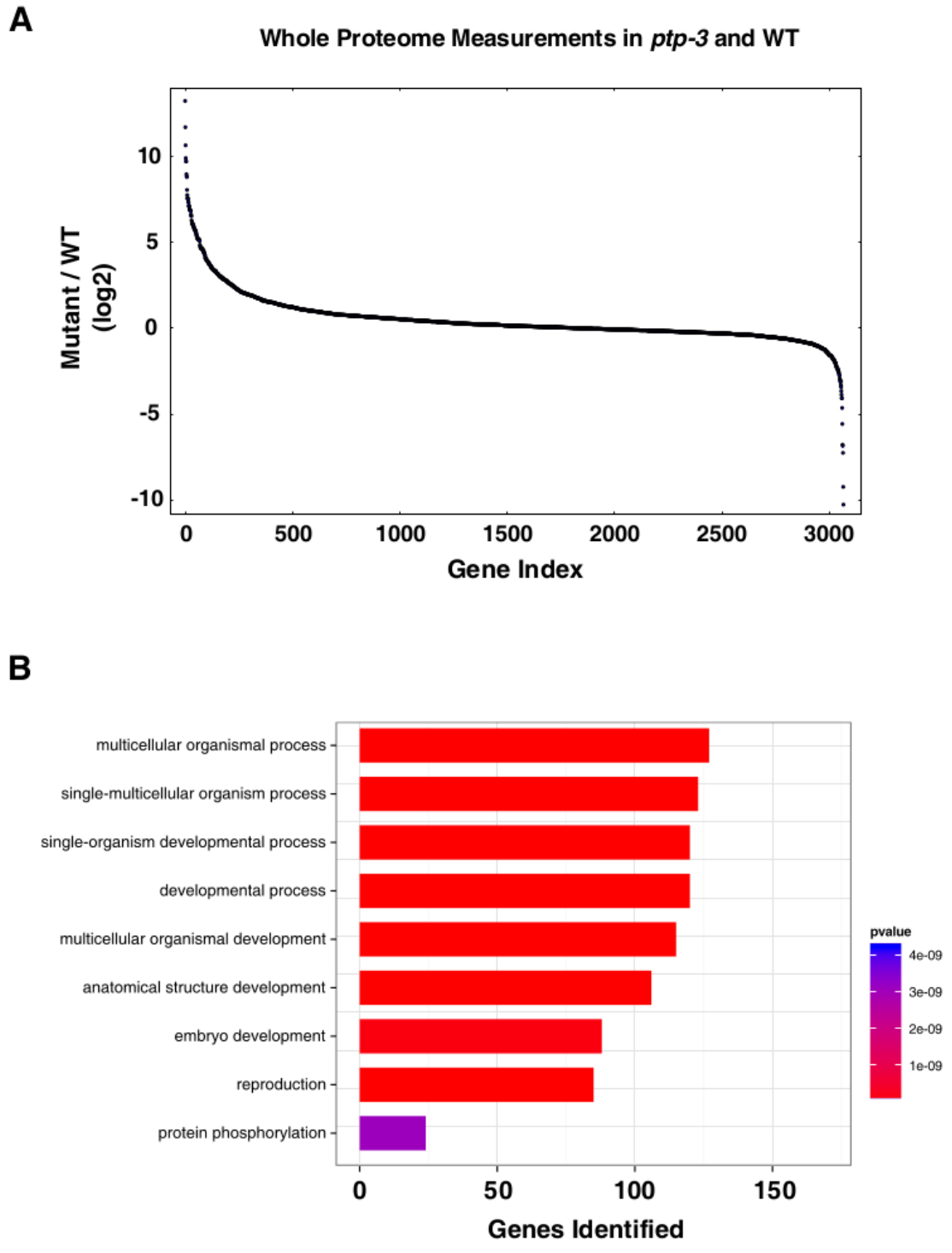
C. Elegans Protein (Human Gene Symbol)	Protein Domains	Sequence	Mutant/Wildtype Fold Change
MAPK15		YPEGQKMPDLTEYVATRWYRSPEIL	2.38
FER		VSDFGLSRVGTEYQMDPSKRVPIRW	2.98
MAPK7		RDDANVGGHMTQYVSTRWYRAPEIL	2.89
NTRK2		IRISDFGMSRRLYDHSEYYTMDHRG	6.81
HIPK1		SASHRSKAVTNTYLQSRYYRAPEII	2.74
GSK3A		SAIESVKTPQQSYHVTRYYRPPELL	2.13
CDK5		YDKMEKIGEGTYGTVFKARNKNSG	4.68
MET		GYLSVESQGGPKYTQLTMQDSKETA	1.88
PDGFRB		TQKWSHFASANNYMDIQALANANKK	>10

Figure 4



Supplementary Figure 1



CHAPTER 5

Concluding Remarks

In my thesis, I set out to characterize the substrates of a tyrosine phosphatase through SILAC based phosphoproteomics. Because we used an organism instead of cell culture, metabolic pathways interfered with our labeling strategy and resulted in data that was not compatible with existing programs. To rectify this, PyQuant was created and used to quantify the *C. elegans* data. This analysis revealed a plethora of putative substrates of the *C. elegans* protein tyrosine phosphatase *ptp-3* and uncovered the few known substrates of the human orthologs of *ptp-3*. In total, there are approximately 9 known substrates of the LAR family. Following this study, we reported 255 putative substrates of the LAR family, which is over an order of magnitude increase in the possible substrates of the LAR family. Though these are all putative substrates and need additional validation measures, the conservation of many putative substrates from *C. elegans* to humans and the number of known substrates rediscovered in the *C. elegans* data suggest a fair portion of the putative substrates may validate. In addition to the work described herein, we attempted two validation measures: shRNA knockdown of *ptp-3* and a GST fusion protein. These validation efforts were unfortunately unsuccessful. The shRNA did not completely knock down the phosphatase as it was still found in the mass spectrometry data and the GST fusion protein was functionally inactive. Thus, the proper folding conditions need to be established to validate these substrates in *C. elegans* or a GST fusion protein from another species may need to be used in its stead.

Following the use of PyQuant to handle the *C. elegans* data, it was further developed to solve an assortment of problems in quantitative mass spectrometry. A notable case is the "missing value problem" in proteomics. A recurrent issue in replicate analysis, especially in phosphotyrosine enrichment assays, is poor concordance between samples. Due to DDA, the notion of a replicate in mass spectrometry is different than in most experimental assays. For instance, a western blot does not randomly sample the proteins being transferred in an all-or-none fashion. However, most experimentalists have the same expectations of a mass spectrometry replicate as a western blot replicate. This "missing value analysis" feature will significantly help in complex samples such as phosphotyrosine enrichment, where significant improvements in replicate concordance have already been observed.

Finally, a central ideology of PyQuant is that a user should be able to interrogate their data and see exactly how it was quantified. In existing platforms, the chain of custody of how data is manipulated and quantified is obscured. Proteome Discoverer currently gives the most information by reporting the abundance of each isotopic cluster. However – the XIC, the isotopes chosen in each scan, and how the data was fit – the factors that comprise the final data are hidden. Whatever their reasons are for not presenting this data, it can mask flaws in the data analysis. Given that in a normal proteomics experiment, there may only be tens or hundreds of peptides of interest, manual validation on every peptide is not a cumbersome routine as it is exceedingly easy to visually determine if data is trustworthy. Thus, while existing applications ask the user

to trust their quantification of ions, PyQuant lets the user manually validate ions that may be extremely important for future experimental directions.

CHAPTER 6

References

Alves, G., Wu, W.W., Wang, G., Shen, R.-F., Yu, Y.-K., 2008. Enhancing Peptide Identification Confidence by Combining Search Methods. J. Proteome Res. 7, 3102–3113. doi:10.1021/pr700798h

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G., 2000. Gene Ontology: tool for the unification of biology. Nat. Genet. 25, 25–29. doi:10.1038/75556

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S., 2009. MEME Suite: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208. doi:10.1093/nar/gkp335

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. ISMB Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36.

Bard-Chapeau, E.A., Li, S., Ding, J., Zhang, S.S., Zhu, H.H., Princen, F., Fang, D.D., Han, T., Bailly-Maitre, B., Poli, V., Varki, N.M., Wang, H., Feng, G.-S., 2011. Ptpn11/Shp2 Acts as a Tumor Suppressor in Hepatocellular Carcinogenesis. Cancer Cell 19, 629–639. doi:10.1016/j.ccr.2011.03.023

Baxter, R.M., Secrist, J.P., Vaillancourt, R.R., Kazlauskas, A., 1998. Full Activation of the Platelet-derived Growth Factor β-Receptor Kinase Involves Multiple Events. J. Biol. Chem. 273, 17050–17055. doi:10.1074/jbc.273.27.17050

Bicho, C.C., Alves, F. de L., Chen, Z.A., Rappsilber, J., Sawin, K.E., 2010. A Genetic Engineering Solution to the "Arginine Conversion Problem" in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). Mol. Cell. Proteomics 9, 1567–1577. doi:10.1074/mcp.M110.000208

Borek, W.E., Zou, J., Rappsilber, J., Sawin, K.E., 2015. Deletion of Genes Encoding Arginase Improves Use of "Heavy" Isotope-Labeled Arginine for Mass Spectrometry in Fission Yeast. PLoS ONE 10, e0129548. doi:10.1371/journal.pone.0129548

Buchdunger, E., Matter, A., Druker, B.J., 2001. Bcr-Abl inhibition as a modality of CML therapeutics. Biochim. Biophys. Acta BBA - Rev. Cancer 1551, M11–M18. doi:10.1016/S0304-419X(01)00022-1

Chaerkady, R., Kelkar, D.S., Muthusamy, B., Kandasamy, K., Dwivedi, S.B., Sahasrabuddhe, N.A., Kim, M.-S., Renuse, S., Pinto, S.M., Sharma, R., Pawar, H., Sekhar, N.R., Mohanty, A.K., Getnet, D., Yang, Y., Zhong, J., Dash, A.P., MacCallum, R.M., Delanghe, B., Mlambo, G., Kumar, A., Prasad, T.S.K., Okulate, M., Kumar, N., Pandey, A., 2011. A proteogenomic analysis of Anopheles gambiae using high-resolution Fourier transform mass spectrometry. Genome Res. 21, 1872–1881. doi:10.1101/gr.127951.111

Chan, G., Kalaitzidis, D., Neel, B.G., 2008. The tyrosine phosphatase Shp2 (PTPN11) in cancer. Cancer Metastasis Rev. 27, 179–192. doi:10.1007/s10555-008-9126-y

Chen, Y.-W., Guo, T., Shen, L., Wong, K.-Y., Tao, Q., Choi, W.W.L., Au-Yeung, R.K.H., Chan, Y.-P., Wong, M.L.Y., Tang, J.C.O., Liu, W.-P., Li, G.-D., Shimizu, N., Loong, F., Tse, E., Kwong, Y.-L., Srivastava, G., 2015. Receptor-type tyrosine-protein phosphatase κ directly targets STAT3 activation for tumor suppression in nasal NK/T-cell lymphoma. Blood 125, 1589–1600. doi:10.1182/blood-2014-07-588970

Chick, J.M., Kolippakkam, D., Nusinow, D.P., Zhai, B., Rad, R., Huttlin, E.L., Gygi, S.P., 2015. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. 33, 743–749. doi:10.1038/nbt.3267

Chin-Sang, I.D., Moseley, S.L., Ding, M., Harrington, R.J., George, S.E., Chisholm, A.D., 2002. The divergent C. elegans ephrin EFN-4 functions inembryonic morphogenesis in a pathway independent of the VAB-1 Eph receptor. Development 129, 5499–5510. doi:10.1242/dev.00122

Cox, J., Mann, M., 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. 26, 1367–1372. doi:10.1038/nbt.1511

Craig, R., Beavis, R.C., 2004. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20, 1466–1467. doi:10.1093/bioinformatics/bth092

Dekel, B., Biton, S., Yerushalmi, G.M., Altstock, R.T., Mittelman, L., Faletto, D., Smordinski, N.I., Tsarfaty, I., 2003. In situ activation pattern of Met docking site following renal injury and hypertrophy. Nephrol. Dial. Transplant. 18, 1493–1504. doi:10.1093/ndt/gfg215

Deutsch, E.W., Mendoza, L., Shteynberg, D., Farrah, T., Lam, H., Tasman, N., Sun, Z., Nilsson, E., Pratt, B., Prazen, B., Eng, J.K., Martin, D.B., Nesvizhskii, A.I., Aebersold, R., 2010. A guided tour of the Trans-Proteomic Pipeline. PROTEOMICS 10, 1150–1159. doi:10.1002/pmic.200900375

Duan, G., Li, X., Köhn, M., 2015. The human DEPhOsphorylation database DEPOD: a 2015 update. Nucleic Acids Res. 43, D531–D535. doi:10.1093/nar/gku1009

Eng, J.K., Jahan, T.A., Hoopmann, M.R., 2013. Comet: An open-source MS/MS sequence database search tool. PROTEOMICS 13, 22–24. doi:10.1002/pmic.201200439

Eng, J.K., McCormack, A.L., Yates, J.R., 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J. Am. Soc. Mass Spectrom. 5, 976–989. doi:10.1016/1044-0305(94)80016-2

Faux, C., Hawadle, M., Nixon, J., Wallace, A., Lee, S., Murray, S., Stoker, A., 2007. PTPσ binds and dephosphorylates neurotrophin receptors and can suppress NGF-dependent neurite outgrowth from sensory neurons. Biochim. Biophys. Acta BBA - Mol. Cell Res. 1773, 1689–1700. doi:10.1016/j.bbamcr.2007.06.008

Fredens, J., Engholm-Keller, K., Giessing, A., Pultz, D., Larsen, M.R., Højrup, P., Møller-Jensen, J., Færgeman, N.J., 2011. Quantitative proteomics by amino acid labeling in C. elegans. Nat. Methods 8, 845–847. doi:10.1038/nmeth.1675

Frith, M.C., Saunders, N.F.W., Kobe, B., Bailey, T.L., 2008. Discovering Sequence Motifs with Arbitrary Insertions and Deletions. PLoS Comput Biol 4, e1000071. doi:10.1371/journal.pcbi.1000071

FUNATO, K., YAMAZUMI, Y., ODA, T., AKIYAMA, T., 2011. Tyrosine phosphatase PTPRD suppresses colon cancer cell migration in coordination with CD44. Exp. Ther. Med. 2, 457–463. doi:10.3892/etm.2011.231

Giefing, M., Zemke, N., Brauze, D., Kostrzewska-Poczekaj, M., Luczak, M., Szaumkessel, M., Pelinska, K., Kiwerska, K., Tönnies, H., Grenman, R., Figlerowicz, M., Siebert, R., Szyfter, K., Jarmuz, M., 2011. High resolution ArrayCGH and expression profiling identifies PTPRD and PCDH17/PCH68 as tumor suppressor gene candidates in laryngeal squamous cell carcinoma. Genes. Chromosomes Cancer 50, 154–166. doi:10.1002/gcc.20840

Grant, C.E., Bailey, T.L., Noble, W.S., 2011. FIMO: scanning for occurrences of a given motif. Bioinformatics 27, 1017–1018. doi:10.1093/bioinformatics/btr064

Hanahan, D., Weinberg, R.A., 2000. The Hallmarks of Cancer. Cell 100, 57–70. doi:10.1016/S0092-8674(00)81683-9

Harrington, R.J., Gutch, M.J., Hengartner, M.O., Tonks, N.K., Chisholm, A.D., 2002. The C. elegans LAR-like receptor tyrosine phosphatase PTP-3 and the VAB-1 Eph

receptor tyrosine kinase have partly redundant functions in morphogenesis. Development 129, 2141–2153.

Harris, T.W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W.J., Cruz, N.D.L., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Müller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers, A., Schindelman, G., Schwarz, E.M., Tuli, M.A., Auken, K.V., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L.D., Spieth, J., Sternberg, P.W., 2010. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 38, D463–D467. doi:10.1093/nar/gkp952

Hashimoto, N., Feener, E.P., Zhang, W.R., Goldstein, B.J., 1992. Insulin receptor protein-tyrosine phosphatases. Leukocyte common antigen-related phosphatase rapidly deactivates the insulin receptor kinase by preferential dephosphorylation of the receptor regulatory domain. J. Biol. Chem. 267, 13811–13814.

Hebert, A.S., Merrill, A.E., Bailey, D.J., Still, A.J., Westphall, M.S., Strieter, E.R., Pagliarini, D.J., Coon, J.J., 2013. Neutron-encoded mass signatures for multiplexed proteome quantification. Nat. Methods 10, 332–334. doi:10.1038/nmeth.2378

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., Skrzypek, E., 2015. PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. 43, D512–D520. doi:10.1093/nar/gku1267

Hultin-Rosenberg, L., Forshed, J., Branca, R.M.M., Lehtiö, J., Johansson, H.J., 2013. Defining, Comparing, and Improving iTRAQ Quantification in Mass

Spectrometry Proteomics Data. Mol. Cell. Proteomics 12, 2021–2031. doi:10.1074/mcp.M112.021592

Jacob, S.T., Motiwala, T., 2005. Epigenetic regulation of protein tyrosine phosphatases: potential molecular targets for cancer therapy. Cancer Gene Ther. 12, 665–672. doi:10.1038/sj.cgt.7700828

Kim, M.-S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., Thomas, J.K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N.A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L.D.N., Patil, A.H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S.K., Marimuthu, A., Sathe, G.J., Chavan, S., Datta, K.K., Subbannayya, Y., Sahu, A., Yelamanchi, S.D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K.R., Syed, N., Goel, R., Khan, A.A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P.G., Freed, D., Zahari, M.S., Mukherjee, K.K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C.J., Shankar, S.K., Satishchandra, P., Schroeder, J.T., Sirdeshmukh, R., Maitra, A., Leach, S.D., Drake, C.G., Halushka, M.K., Prasad, T.S.K., Hruban, R.H., Kerr, C.L., Bader, G.D., Iacobuzio-Donahue, C.A., Gowda, H., Pandey, A., 2014a. A draft map of the human proteome. Nature 509, 575–581. doi:10.1038/nature13302

Kim, M.-S., Zhong, Y., Yachida, S., Rajeshkumar, N.V., Abel, M.L., Marimuthu, A., Mudgal, K., Hruban, R.H., Poling, J.S., Tyner, J.W., Maitra, A., Iacobuzio-Donahue, C.A., Pandey, A., 2014b. Heterogeneity of Pancreatic Cancer

Metastases in a Single Patient Revealed by Quantitative Proteomics. Mol. Cell. Proteomics 13, 2803–2811. doi:10.1074/mcp.M114.038547

Kolibaba, K.S., Druker, B.J., 1997. Protein tyrosine kinases and cancer. Biochim. Biophys. Acta BBA - Rev. Cancer 1333, F217–F248. doi:10.1016/S0304-419X(97)00022-X

Kwon, T., Choi, H., Vogel, C., Nesvizhskii, A.I., Marcotte, E.M., 2011. MSblender: A Probabilistic Approach for Integrating Peptide Identifications from Multiple Database Search Engines. J. Proteome Res. 10, 2949–2958. doi:10.1021/pr2002116

Larance, M., Bailly, A.P., Pourkarimi, E., Hay, R.T., Buchanan, G., Coulthurst, S., Xirodimas, D.P., Gartner, A., Lamond, A.I., 2011. Stable Isotope Labeling with Amino acids in Nematodes. Nat. Methods 8, 849–851. doi:10.1038/nmeth.1679

Lasonder, E., Ishihama, Y., Andersen, J.S., Vermunt, A.M.W., Pain, A., Sauerwein, R.W., Eling, W.M.C., Hall, N., Waters, A.P., Stunnenberg, H.G., Mann, M., 2002. Analysis of the Plasmodium falciparum proteome by high-accuracy mass spectrometry. Nature 419, 537–542. doi:10.1038/nature01111

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R., Bigner, S.H., Giovanella, B.C., Ittmann, M., Tycko, B., Hibshoosh, H., Wigler, M.H., Parsons, R., 1997. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 275, 1943–1947.

Ma, B., Zhang, K., Hendrie, C., Liang, C., Li, M., Doherty-Kirby, A., Lajoie, G., 2003. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. RCM 17, 2337–2342. doi:10.1002/rcm.1196

Machide, M., Hashigasako, A., Matsumoto, K., Nakamura, T., 2006. Contact Inhibition of Hepatocyte Growth Regulated by Functional Association of the c-Met/Hepatocyte Growth Factor Receptor and LAR Protein-tyrosine Phosphatase. J. Biol. Chem. 281, 8765–8772. doi:10.1074/jbc.M512298200

Merrill, A.E., Hebert, A.S., MacGilvray, M.E., Rose, C.M., Bailey, D.J., Bradley, J.C., Wood, W.W., Masri, M.E., Westphall, M.S., Gasch, A.P., Coon, J.J., 2014. NeuCode Labels for Relative Protein Quantification. Mol. Cell. Proteomics 13, 2503–2512. doi:10.1074/mcp.M114.040287

Mi, H., Muruganujan, A., Casagrande, J.T., Thomas, P.D., 2013. Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. 8, 1551–1566. doi:10.1038/nprot.2013.092

Mitchell, C.J., Getnet, D., Kim, M.-S., Manda, S.S., Kumar, P., Huang, T.-C., Pinto, S.M., Nirujogi, R.S., Iwasaki, M., Shaw, P.G., Wu, X., Zhong, J., Chaerkady, R., Marimuthu, A., Muthusamy, B., Sahasrabuddhe, N.A., Raju, R., Bowman, C., Danilova, L., Cutler, J., Kelkar, D.S., Drake, C.G., Prasad, T.S., Marchionni, L., Murakami, P.N., Scott, A.F., Shi, L., Thierry-Mieg, J., Thierry-Mieg, D., Irizarry, R., Cope, L., Ishihama, Y., Wang, C., Gowda, H., Pandey, A., 2015. A multi-omic

analysis of human naïve CD4+ T cells. BMC Syst. Biol. 9, 75. doi:10.1186/s12918-015-0225-4

Morris, L.G.T., Taylor, B.S., Bivona, T.G., Gong, Y., Eng, S., Brennan, C.W., Kaufman, A., Kastenhuber, E.R., Banuchi, V.E., Singh, B., Heguy, A., Viale, A., Mellinghoff, I.K., Huse, J., Ganly, I., Chan, T.A., 2011. Genomic dissection of the epidermal growth factor receptor (EGFR)/PI3K pathway reveals frequent deletion of the EGFR phosphatase PTPRS in head and neck cancers. Proc. Natl. Acad. Sci. U. S. A. 108, 19024–19029. doi:10.1073/pnas.1111963108

Müller, T., Choidas, A., Reichmann, E., Ullrich, A., 1999. Phosphorylation and Free Pool of β-Catenin Are Regulated by Tyrosine Kinases and Tyrosine Phosphatases during Epithelial Cell Migration. J. Biol. Chem. 274, 10173–10183. doi:10.1074/jbc.274.15.10173

Murphy, J.P., Stepanova, E., Everley, R.A., Paulo, J.A., Gygi, S.P., 2015. Comprehensive Temporal Protein Dynamics during the Diauxic Shift in Saccharomyces cerevisiae. Mol. Cell. Proteomics 14, 2454–2465. doi:10.1074/mcp.M114.045849

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D.B., Steen, H., Pandey, A., Mann, M., 2002. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. Mol. Cell. Proteomics 1, 376–386. doi:10.1074/mcp.M200025-MCP200

Ong, S.-E., Kratchmarova, I., Mann, M., 2003. Properties of 13C-Substituted Arginine in Stable Isotope Labeling by Amino Acids in Cell Culture (SILAC). J. Proteome Res. 2, 173–181. doi:10.1021/pr0255708

Ow, S.Y., Salim, M., Noirel, J., Evans, C., Rehman, I., Wright, P.C., 2009. iTRAQ Underestimation in Simple and Complex Mixtures: "The Good, the Bad and the Ugly." J. Proteome Res. 8, 5347–5355. doi:10.1021/pr900634c

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999a. Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS 20, 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Perkins, D.N., Pappin, D.J.C., Creasy, D.M., Cottrell, J.S., 1999b. Probability-based protein identification by searching sequence databases using mass spectrometry data. ELECTROPHORESIS 20, 3551–3567. doi:10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Pinto, S.M., Nirujogi, R.S., Rojas, P.L., Patil, A.H., Manda, S.S., Subbannayya, Y., Roa, J.C., Chatterjee, A., Prasad, T.S.K., Pandey, A., 2015. Quantitative phosphoproteomic analysis of IL-33-mediated signaling. Proteomics 15, 532–544. doi:10.1002/pmic.201400303

Purvine, S., Eppel*, J.-T., Yi, E.C., Goodlett, D.R., 2003. Shotgun collision-induced dissociation of peptides using a time of flight mass analyzer. PROTEOMICS 3, 847–850. doi:10.1002/pmic.200300362

Renuse, S., Chaerkady, R., Pandey, A., 2011. Proteogenomics. PROTEOMICS 11, 620–630. doi:10.1002/pmic.201000615

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin,

S., Bartlet-Jones, M., He, F., Jacobson, A., Pappin, D.J., 2004. Multiplexed Protein Quantitation in Saccharomyces cerevisiae Using Amine-reactive Isobaric Tagging Reagents. Mol. Cell. Proteomics 3, 1154–1169. doi:10.1074/mcp.M400129-MCP200

Rousseeuw, P.J., Driessen, K.V., 1999. A Fast Algorithm for the Minimum Covariance Determinant Estimator. Technometrics 41, 212–223. doi:10.1080/00401706.1999.10485670

Rush, J., Moritz, A., Lee, K.A., Guo, A., Goss, V.L., Spek, E.J., Zhang, H., Zha, X.-M., Polakiewicz, R.D., Comb, M.J., 2005. Immunoaffinity profiling of tyrosine phosphorylation in cancer cells. Nat. Biotechnol. 23, 94–101. doi:10.1038/nbt1046

Schwartz, D., Gygi, S.P., 2005. An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. Nat. Biotechnol. 23, 1391–1398. doi:10.1038/nbt1146

Schwarz, G., 1978. Estimating the Dimension of a Model. Ann. Stat. 6, 461–464. doi:10.1214/aos/1176344136

Serang, O., Noble, W., 2012. A review of statistical methods for protein identification using tandem mass spectrometry. Stat. Interface 5, 3–20.

Shteynberg, D., Deutsch, E.W., Lam, H., Eng, J.K., Sun, Z., Tasman, N., Mendoza, L., Moritz, R.L., Aebersold, R., Nesvizhskii, A.I., 2011. iProphet: Multi-level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein

Identification Rates and Error Estimates. Mol. Cell. Proteomics 10, M111.007690. doi:10.1074/mcp.M111.007690

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J.D., Higgins, D.G., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539. doi:10.1038/msb.2011.75

Simonis, N., Rual, J.-F., Carvunis, A.-R., Tasan, M., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Sahalie, J.M., Venkatesan, K., Gebreab, F., Cevik, S., Klitgord, N., Fan, C., Braun, P., Li, N., Ayivi-Guedehoussou, N., Dann, E., Bertin, N., Szeto, D., Dricot, A., Yildirim, M.A., Lin, C., de Smet, A.-S., Kao, H.-L., Simon, C., Smolyar, A., Ahn, J.S., Tewari, M., Boxem, M., Milstein, S., Yu, H., Dreze, M., Vandenhaute, J., Gunsalus, K.C., Cusick, M.E., Hill, D.E., Tavernier, J., Roth, F.P., Vidal, M., 2009. Empirically-controlled mapping of the Caenorhabditis elegans protein-protein interactome network. Nat. Methods 6, 47–54.

Slamon, D.J., Clark, G.M., Wong, S.G., Levin, W.J., Ullrich, A., McGuire, W.L., 1987. Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/neu Oncogene. Science, New Series 235, 177–182.

Slebos, R.J.C., Wang, X., Wang, X., Zhang, B., Tabb, D.L., Liebler, D.C., 2015. Proteomic analysis of colon and rectal carcinoma using standard and customized databases. Sci. Data 2, 150022. doi:10.1038/sdata.2015.22

Stiernagle, T., 2006. Maintenance of C. elegans. Wormbook. doi:10.1895/wormbook.1.101.1

Sury, M.D., Chen, J.-X., Selbach, M., 2010. The SILAC Fly Allows for Accurate Protein Quantification in Vivo. Mol. Cell. Proteomics 9, 2173–2183. doi:10.1074/mcp.M110.000323

Syed, N., Barbhuiya, M.A., Pinto, S.M., Nirujogi, R.S., Renuse, S., Datta, K.K., Khan, A.A., Srikumar, K., Prasad, T.S.K., Kumar, M.V., Kumar, R.V., Chatterjee, A., Pandey, A., Gowda, H., 2015. Phosphotyrosine profiling identifies ephrin receptor A2 as a potential therapeutic target in esophageal squamous-cell carcinoma. PROTEOMICS 15, 374–382. doi:10.1002/pmic.201400379

Tan, C.S.H., Bodenmiller, B., Pasculescu, A., Jovanovic, M., Hengartner, M.O., Jørgensen, C., Bader, G.D., Aebersold, R., Pawson, T., Linding, R., 2009. Comparative analysis reveals conserved protein phosphorylation networks implicated in multiple diseases. Sci. Signal. 2, ra39. doi:10.1126/scisignal.2000316

Taylor, J.A., Johnson, R.S., 1997. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. Rapid Commun. Mass Spectrom. RCM 11, 1067–1075. doi:10.1002/(SICI)1097-0231(19970615)11:9<1067::AID-RCM953>3.0.CO;2-L

Tharakan, R., Edwards, N., Graham, D.R.M., 2010. Data maximization by multipass analysis of protein mass spectra. PROTEOMICS 10, 1160–1171. doi:10.1002/pmic.200900433

The Cancer Genome Atlas Network, 2012. Comprehensive molecular portraits of human breast tumours. Nature 490, 61–70. doi:10.1038/nature11412

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Hamon, C., 2003. Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. Anal. Chem. 75, 1895–1904. doi:10.1021/ac0262560

Ting, L., Rad, R., Gygi, S.P., Haas, W., 2011. MS3 eliminates ratio distortion in isobaric labeling-based multiplexed quantitative proteomics. Nat. Methods 8, 937–940. doi:10.1038/nmeth.1714

Tonks, N.K., 2006. Protein tyrosine phosphatases: from genes, to function, to disease. Nat. Rev. Mol. Cell Biol. 7, 833–846. doi:10.1038/nrm2039

Uetani, N., Chagnon, M.J., Kennedy, T.E., Iwakura, Y., Tremblay, M.L., 2006. Mammalian Motoneuron Axon Targeting Requires Receptor Protein Tyrosine Phosphatases σ and δ. J. Neurosci. 26, 5872–5880. doi:10.1523/JNEUROSCI.0386-06.2006

Van Hoof, D., Pinkse, M.W.H., Oostwaard, D.W.-V., Mummery, C.L., Heck, A.J.R., Krijgsveld, J., 2007. An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics. Nat. Methods 4, 677–678. doi:10.1038/nmeth0907-677

Veeriah, S., Brennan, C., Meng, S., Singh, B., Fagin, J.A., Solit, D.B., Paty, P.B., Rohle, D., Vivanco, I., Chmielecki, J., Pao, W., Ladanyi, M., Gerald, W.L., Liau, L., Cloughesy, T.C., Mischel, P.S., Sander, C., Taylor, B., Schultz, N., Major, J., Heguy, A., Fang, F., Mellinghoff, I.K., Chan, T.A., 2009. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and

mutated in glioblastoma and other human cancers. Proc. Natl. Acad. Sci. 106, 9435–9440. doi:10.1073/pnas.0900571106

Vijayvargia, R., Kaur, S., Krishnasastry, M.V., 2004. α-Hemolysin-induced dephosphorylation of EGF receptor of A431 cells is carried out by rPTPσ. Biochem. Biophys. Res. Commun. 325, 344–352. doi:10.1016/j.bbrc.2004.10.038

Wang, X., Slebos, R.J.C., Wang, D., Halvey, P.J., Tabb, D.L., Liebler, D.C., Zhang, B., 2012. Protein identification using customized protein sequence databases derived from RNA-Seq data. J. Proteome Res. 11, 1009–1017. doi:10.1021/pr200766z

Washburn, M.P., Ulaszek, R., Deciu, C., Schieltz, D.M., Yates, J.R., 2002. Analysis of quantitative proteomic data generated via multidimensional protein identification technology. Anal. Chem. 74, 1650–1657.

Wu, L., Candille, S.I., Choi, Y., Xie, D., Jiang, L., Li-Pook-Than, J., Tang, H., Snyder, M., 2013. Variation and genetic control of protein abundance in humans. Nature 499, 79–82. doi:10.1038/nature12223

Yu, G., Wang, L.-G., Han, Y., He, Q.-Y., 2012. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. OMICS J. Integr. Biol. 16, 284–287. doi:10.1089/omi.2011.0118

Zahari, M.S., Wu, X., Blair, B.G., Pinto, S.M., Nirujogi, R.S., Jelinek, C.A., Malhotra, R., Kim, M.-S., Park, B.H., Pandey, A., 2015. Activating Mutations in PIK3CA Lead to Widespread Modulation of the Tyrosine Phosphoproteome. J. Proteome Res. 14, 3882–3891. doi:10.1021/acs.jproteome.5b00302

Zanivan, S., Krueger, M., Mann, M., 2012. In Vivo Quantitative Proteomics: The SILAC Mouse, in: Shimaoka, M. (Ed.), Integrin and Cell Adhesion Molecules, Methods in Molecular Biology. Humana Press, pp. 435–450.

Zhong, J., Kim, M.-S., Chaerkady, R., Wu, X., Huang, T.-C., Getnet, D., Mitchell, C.J., Palapetta, S.M., Sharma, J., O'Meally, R.N., Cole, R.N., Yoda, A., Moritz, A., Loriaux, M.M., Rush, J., Weinstock, D.M., Tyner, J.W., Pandey, A., 2012. TSLP Signaling Network Revealed by SILAC-Based Phosphoproteomics. Mol. Cell. Proteomics 11, M112.017764. doi:10.1074/mcp.M112.017764

Zukerberg, L.R., Patrick, G.N., Nikolic, M., Humbert, S., Wu, C.-L., Lanier, L.M., Gertler, F.B., Vidal, M., Van Etten, R.A., Tsai, L.-H., 2000. Cables Links Cdk5 and c-Abl and Facilitates Cdk5 Tyrosine Phosphorylation, Kinase Upregulation, and Neurite Outgrowth. Neuron 26, 633–646. doi:10.1016/S0896-6273(00)81200-3

Curriculum Vita

Christopher J. Mitchell

ADDRESS 1612 Patapsco St.

Baltimore, Maryland, USA 21230

TELEPHONE (609)-425-3536

E-MAIL cmitch48@jhmi.edu

EDUCATION

Doctor of Philosophy, 2015

Department of Biochemistry, Cellular and Molecular Biology, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

Bachelor of Arts, 2010

Biology, University of Pennsylvania, Philadelphia, Pennsylvania, U.S.A

PUBLICATIONS (peer-reviewed original articles)

1. **Christopher Mitchell***, Derese Getnet*, Min-Sik Kim*, Srinivas Srikanth, Praveen Kumar, Sneha Pinto, Mio Iwasaki, Tai-Chung Huang, Patrick Shaw, Xinyan Wu, Jun Zhong, Raghothama Chaerkady, Babylakshmi Muthusamy, Raja Sekhar Nirujogi, Nandini Sahasrabuddhe, Rajesh Raju, Caitlyn Bowman, Ludmila Danilova, Jevon Cutler, Dhanashree Kelkar, Charles G. Drake, T. S. Keshava Prasad, Luigi

Marchionni, Peter Marukami, Alan Scott, Leming Shi, Jean Thierry-Mieg, Danielle Thierry-Mieg, Rafael Irizarry, Charles Wang, Leslie Cope, Yasushi Ishihama, H. C. Harsha Gowda, and Akhilesh Pandey. (2015). **A Multi-Omic Analysis of Human Naïve CD4+ T cells.** *BMC Systems Biology.* *Shared first-authorship.

2. Baras, A.S.*, **Mitchell, C.J*.**, Myers, J.R., Gupta, S., Weng, L., Ashton, J.M., Cornish, T.C., Pandey, A., Halushka, M.K. (2015). **miRge - A Multiplexed Method of Processing Small RNA-Seq Data to Determine MicroRNA Entropy. PLOS ONE.** **Shared first-authorship.*
3. Kim, M. S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T. C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., **Mitchell, C. J**., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. and Pandey, A. (2014). **A draft map of the human proteome**. *Nature.*

4. Kelkar, D. S., Provost. E., Chaerkady, R., Muthusamy, B., Manda, S. S., Subbannayya, T., Selvan, L. D. N., Wang, C. H., Datta, K. K., Woo, S., Dwivedi, S. B., Renuse, S., Getnet, D., Huang, T. C., Kim, M. S., Pinto, S. M., **Mitchell, C. J.**, Madugundu, A. K., Kumar, P., Sharma, J. Advani, J., Dey, G., Balakrishnan, L., Syed, N., Nanjappa, V., Subbannayya, Y., Goel, R., Prasad, T. S. K., Bafna, V., Sirdeshmukh, R., Gowda, H., Wang, C., Leach, S. D. and Pandey, A. (2014). **Annotation of the zebrafish genome through an integrated transcriptomic and proteomic analysis.** Molecular and Cellular Proteomics.
5. Heydarian, M., Luperchio, T. R., Cutler, J., **Mitchell, C. J.**, Kim, M. S., Pandey, A., Sollner-Webb, B. and Reddy, K. (2014). **Prediction of gene activity in early B cell development based on an integrative multi-Omics Analysis.** Journal of Proteomics and Bioinformatics.
6. Zhong, J., Kim, M. S., Chaerkady, R., Wu, X., Huang, T. C., Getnet, D., **Mitchell, C. J.**, Palapetta, S. M., Sharma, J., O'Meally, R. N., Cole, R. N., Yoda, A., Moritz, A., Loriaux, M. M., Rush, J., Weinstock, D. M., Tyner, J. W. and Pandey, A. (2012). **TSLP signaling network revealed by SILAC-based phosphoproteomics.** Molecular and Cellular Proteomics.
7. Schroeder HW 3rd, **Mitchell C.**, Shuman H, Holzbaur EL, Goldman YE. (2010). **Motor Number Controls Cargo Switching at Actin-Microtubule Intersections In Vitro.** *Current Biology*